

1. [Preface](#)
2. Sampling and Data
 1. [Introduction](#)
 2. [Definitions of Statistics, Probability, and Key Terms](#)
3. Descriptive Statistics
 1. [Introduction](#)
 2. [Measures of the Location of the Data](#)
 3. [Measures of the Spread of the Data](#)
4. Probability Topics
 1. [Introduction](#)
 2. [Terminology](#)
5. Linear Regression and Correlation
 1. [Introduction](#)
 2. [Scatter Plots](#)
 3. [Testing the Significance of the Correlation Coefficient](#)
6. [Review Exercises \(Ch 3-13\)](#)
7. [Practice Tests \(1-4\) and Final Exams](#)
8. [Data Sets](#)
9. [Group and Partner Projects](#)
10. [Solution Sheets](#)
11. [Mathematical Phrases, Symbols, and Formulas](#)
12. [Notes for the TI-83, 83+, 84, 84+ Calculators](#)
13. [Tables](#)

Preface

Introductory Statistics is intended for the one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. In addition to end of section practice and homework sets, examples of each topic are explained step-by-step throughout the text and followed by a Try It problem that is designed as extra practice for students. This book also includes collaborative exercises and statistics labs designed to give students the opportunity to work together and explore key concepts. To support today's student in understanding technology, this book features TI 83, 83+, 84, or 84+ calculator instructions at strategic points throughout. While the book has been built so that each chapter builds on the previous, it can be rearranged to accommodate any instructor's particular needs.

Welcome to *Introductory Statistics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

The foundation of this textbook is *Collaborative Statistics*, by Barbara Illowsky and Susan Dean. Additional topics, examples, and innovations in terminology and practical applications have been added, all with a goal of increasing relevance and accessibility for students.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP[®] courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax's resources

Customization

Introductory Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit your book page on OpenStax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and in low-cost print and iBooks editions.

About *Introductory Statistics*

Introductory Statistics follows scope and sequence requirements of a one-semester introduction to statistics course and is geared toward students majoring in fields other than math or engineering. The text assumes some knowledge of intermediate algebra and focuses on statistics application over theory. *Introductory Statistics* includes innovative practical applications that make the text relevant and accessible, as well as collaborative exercises, technology integration problems, and statistics labs.

Coverage and scope

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 12 Linear Regression and Correlation
Chapter 13 F Distribution and One-Way ANOVA

Alternate sequencing

Introductory Statistics was conceived and written to fit a particular topical sequence, but it can be used flexibly to accommodate other course structures. One such potential structure, which fits reasonably well with the

textbook content, is provided below. Please consider, however, that the chapters were not written to be completely independent, and that the proposed alternate sequence should be carefully considered for student preparation and textual consistency.

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 12 Linear Regression and Correlation
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 13 F Distribution and One-Way ANOVA

Pedagogical foundation and features

Examples are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples about college life and learning, health and medicine, retail and business, and sports and entertainment.

Try It practice problems immediately follow many examples and give students the opportunity to practice as they read the text. **They are usually based on practical and familiar topics, like the Examples themselves.**

Collaborative Exercises provide an in-class scenario for students to work together to explore presented concepts.

Using the TI-83, 83+, 84, 84+ Calculator shows students step-by-step instructions to input problems into their calculator.

The Technology Icon indicates where the use of a TI calculator or computer software is recommended.

Practice, Homework, and Bringing It Together problems give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Statistics labs

These innovative activities were developed by Barbara Illowsky and Susan Dean in order to offer students the experience of designing, implementing, and interpreting statistical analyses. They are drawn from actual experiments and data-gathering processes and offer a unique hands-on and collaborative experience. The labs provide a foundation for further learning and classroom interaction that will produce a meaningful application of statistics.

Statistics Labs appear at the end of each chapter and begin with student learning outcomes, general estimates for time on task, and any global implementation notes. Students are then provided with step-by-step guidance, including sample data tables and calculation prompts. The detailed assistance will help the students successfully apply the concepts in the text and lay the groundwork for future collaborative or individual work.

Additional resources

Student and instructor resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit www.oercommons.org/hubs/OpenStax.

Partner resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on OpenStax.org.

About the authors

Senior contributing authors

Barbara Illowsky, De Anza College
Susan Dean, De Anza College

Contributing authors

Birgit Aquilonius, West Valley College
Charles Ashbacher, Upper Iowa University, Cedar Rapids

Abraham Biggs, Broward Community College
Daniel Birmajer, Nazareth College
Roberta Bloom, De Anza College
Bryan Blount, Kentucky Wesleyan College
Ernest Bonat, Portland Community College
Sarah Boslaugh, Kennesaw State University
David Bosworth, Hutchinson Community College
Sheri Boyd, Rollins College
George Bratton, University of Central Arkansas
Jing Chang, College of Saint Mary
Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Matthew Einsohn, Prescott College
Ann Flanigan, Kapiolani Community College
David French, Tidewater Community College
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Michael Greenwich, College of Southern Nevada
Inna Grushko, De Anza College
Valier Hauber, De Anza College
Janice Hector, De Anza College
Jim Helmreich, Marist College
Robert Henderson, Stephen F. Austin State University
Mel Jacobsen, Snow College
Mary Jo Kane, De Anza College
Lynette Kenyon, Collin County Community College
Charles Klein, De Anza College
Alexander Kolovos
Sheldon Lee, Viterbo University
Sara Lenhart, Christopher Newport University
Wendy Lightheart, Lane Community College
Vladimir Logvenenko, De Anza College
Jim Lucas, De Anza College
Lisa Markus, De Anza College
Miriam Masullo, SUNY Purchase
Diane Mathios, De Anza College
Robert McDevitt, Germanna Community College

Mark Mills, Central College
Cindy Moss, Skyline College
Nydia Nelson, St. Petersburg College
Benjamin Ngwudike, Jackson State University
Jonathan Oaks, Macomb Community College
Carol Olmstead, De Anza College
Adam Pennell, Greensboro College
Kathy Plum, De Anza College
Lisa Rosenberg, Elon University
Sudipta Roy, Kankakee Community College
Javier Rueda, De Anza College
Yvonne Sandoval, Pima Community College
Rupinder Sekhon, De Anza College
Travis Short, St. Petersburg College
Frank Snow, De Anza College
Abdulhamid Sukar, Cameron University
Jeffery Taub, Maine Maritime Academy
Mary Teegarden, San Diego Mesa College
John Thomas, College of Lake County
Philip J. Verrecchia, York College of Pennsylvania
Dennis Walsh, Middle Tennessee State University
Cheryl Wartman, University of Prince Edward Island
Carol Weideman, St. Petersburg College
Andrew Wiesner, Pennsylvania State University

Introduction

class="introduction"

We
encounte
r
statistics
in our
daily
lives
more
often
than we
probably
realize
and from
many
different
sources,
like the
news.
(credit:
David
Sim)



Note:**Chapter Objectives**

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

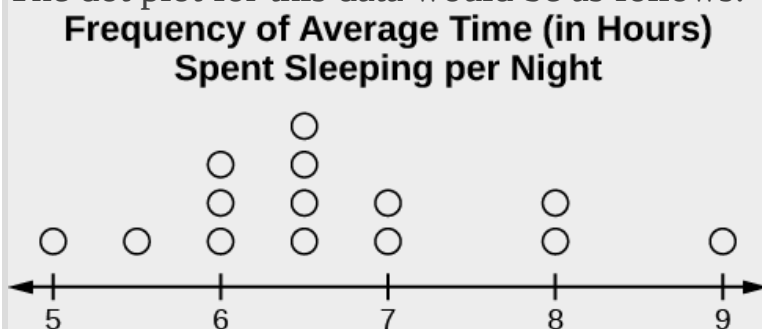
Note:

Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not? Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data.

Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example,

finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or

not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in

inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, usually notated by capital letters such as X and Y , is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:

NOTE

The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:**Exercise:****Problem:**

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

Note:

Try It

Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Solution:

Try It Solutions

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

Example:

Exercise:

Problem:

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population _____ 2. Statistic _____ 3. Parameter _____ 4. Sample _____
5. Variable _____ 6. Data _____

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

Solution:

1. f 2. g 3. e 4. d 5. b 6. c

Example:**Exercise:****Problem:**

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example:**Exercise:****Problem:**

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Note:

Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

The Data and Story Library,
<http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).

Chapter Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12
8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B:

3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12
18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Determine what the key terms refer to in the example for Researcher A.

Exercise:

Problem: population

Solution:

AIDS patients.

Exercise:

Problem: sample

Exercise:

Problem: parameter

Solution:

The average length of time (in months) AIDS patients live after treatment.

Exercise:

Problem: statistic

Exercise:

Problem: variable

Solution:

X = the length of time (in months) AIDS patients live after treatment

HOMEWORK

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

Solution:

- a. all children who take ski or snowboard lessons
- b. a group of these children
- c. the population mean age of children who take their first snowboard lesson
- d. the sample mean age of children who take their first snowboard lesson
- e. X = the age of one child who takes his or her first ski or snowboard lesson
- f. values for X , such as 3, 7, and so on

Exercise:

Problem:

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

Solution:

- a. the clients of the insurance companies
- b. a group of the clients
- c. the mean health costs of the clients
- d. the mean health costs of the sample
- e. X = the health costs of one client
- f. values for X , such as 34, 9, 82, and so on

Exercise:**Problem:**

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Exercise:**Problem:**

A marriage counselor is interested in the proportion of clients she counsels who stay married.

Solution:

- a. all the clients of this counselor
- b. a group of clients of this marriage counselor
- c. the proportion of all her clients who stay married
- d. the proportion of the sample of the counselor's clients who stay married
- e. X = the number of couples who stay married
- f. yes, no

Exercise:**Problem:**

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Exercise:**Problem:**

A marketing company is interested in the proportion of people who will buy a particular product.

Solution:

- a. all people (maybe in a certain geographic area, such as the United States)
- b. a group of the people
- c. the proportion of all people who will buy the product
- d. the proportion of the sample who will buy the product
- e. X = the number of people who will buy it
- f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

- a. all Lake Tahoe Community College students
- b. all Lake Tahoe Community College English students
- c. all Lake Tahoe Community College students in her classes
- d. all Lake Tahoe Community College math students

Exercise:

Problem: Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

- a. variable.
- b. population.
- c. statistic.
- d. data.

Solution:

a

Exercise:

Problem:

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

- a. parameter.
- b. data.
- c. statistic.
- d. variable.

Glossary

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

Introduction

class="introduction"

When you
have large
amounts
of data,
you will
need to
organize
it in a
way that
makes
sense.

These
ballots
from an
election
are rolled
together
with
similar
ballots to
keep them
organized
. (credit:
William
Greeson)

**Note:****Chapter Objectives**

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "**Descriptive Statistics.**" You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Note:

NOTE

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The [Texas Instruments \(TI\) website](#) provides additional instructions for using these calculators.

Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1 , is the same as the 25th percentile, and the third quartile, Q_3 , is the same as the 75th percentile. The median, M , is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

Equation:

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data, and the third quartile, Q_3 , is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, Q_3 , is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The *IQR* can help to determine potential **outliers**. A value is suspected to be a **potential outlier if it is less than $(1.5)(IQR)$ below the first quartile or more than $(1.5)(IQR)$ above the third quartile**. Potential outliers always require further investigation.

Note:**NOTE**

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example:**Exercise:****Problem:**

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000;
387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution:

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000;
575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than $-201,625$. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Note:

Try It

Exercise:**Problem:**

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

\$33,000 \$64,500 \$28,000 \$54,000 \$72,000 \$68,500 \$69,000 \$42,000
\$54,000 \$120,000 \$40,500

Solution:

Order the data from smallest to largest.

\$28,000 \$33,000 \$40,500 \$42,000 \$54,000 \$54,000 \$64,500 \$68,500
\$69,000 \$72,000 \$120,000

Median = \$54,000

$Q_1 = \$40,500$

$Q_3 = \$69,000$

$IQR = \$69,000 - \$40,500 = \$28,500$

$(1.5)(IQR) = (1.5)(\$28,500) = \$42,750$

$Q_1 - (1.5)(IQR) = \$40,500 - \$42,750 = -\$2,250$

$Q_3 + (1.5)(IQR) = \$69,000 + \$42,750 = \$111,750$

No salary is less than $-\$2,250$. However, \$120,000 is more than \$11,750, so \$120,000 is a potential outlier.

Example:**Exercise:**

Problem:

For the two data sets in the [test scores example](#), find the following:

- a. The interquartile range. Compare the two interquartile ranges.
- b. Any outliers in either set.

Solution:

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

- a. The IQR for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The IQR for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

- b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Note:

Try It

Exercise:

Problem:

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class B

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Solution:

Class A

Order the data from smallest to largest.

65 66 67 69 69 76 77 77 79 80 81 83 85 89 90 91 94 96 98 99

$$\text{Median} = \frac{80+81}{2} = 80.5$$

$$Q_1 = \frac{69+76}{2} = 72.5$$

$$Q_3 = \frac{90+91}{2} = 90.5$$

$$IQR = 90.5 - 72.5 = 18$$

Class *B*

Order the data from smallest to largest.

68 68 70 71 72 73 75 78 79 80 80 90 90 92 92 95 95 97 99 100

$$\text{Median} = \frac{80+80}{2} = 80$$

$$Q_1 = \frac{72+73}{2} = 72.5$$

$$Q_3 = \frac{92+95}{2} = 93.5$$

$$IQR = 93.5 - 72.5 = 21$$

The data for Class *B* has a larger *IQR*, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class *B* are more spread out and not clustered about the median.

Example:

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight.** Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3 , is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Note:

Try it

Exercise:**Problem:**

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Solution:

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example:**Exercise:**

Problem: Using [\[link\]](#):

- a. Find the 80th percentile.
- b. Find the 90th percentile.
- c. Find the first quartile. What is another name for the first quartile?

Solution:

Using the data from the frequency table, we have:

- a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
- b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
- c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

Note:

Try It

Exercise:

Problem:

Refer to the [link](#). Find the third quartile. What is another name for the third quartile?

Solution:

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

Note:

Collaborative Statistics

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct two different histograms. For each, starting value = _____ ending value = _____.
4. Find the median, first quartile, and third quartile.
5. Construct a table of the data to find the following:
 - a. the 10th percentile
 - b. the 70th percentile
 - c. the percent of students who own less than four sweaters

A Formula for Finding the k th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n + 1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example:

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- a. Find the 70th percentile.
- b. Find the 83rd percentile.

Solution:

- a.
 - $k = 70$
 - $i = \text{the index}$
 - $n = 29$

$i = \frac{k}{100} (n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

- b.
 - $k = 83^{\text{rd}}$ percentile
 - $i = \text{the index}$
 - $n = 29$

$i = \frac{k}{100} (n + 1) = (\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Note:

Try It

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Calculate the 20th percentile and the 55th percentile.

Solution:

$k = 20$. Index $= i = \frac{k}{100}(n + 1) = \frac{20}{100}(29 + 1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index $= i = \frac{k}{100}(n + 1) = \frac{55}{100}(29 + 1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

Note:**NOTE**

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example:**Exercise:****Problem:**

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- a. Find the percentile for 58.
- b. Find the percentile for 25.

Solution:

- a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

$x = 18$ and $y = 1. \frac{x+0.5y}{n} (100) = \frac{18+0.5(1)}{29} (100) = 63.80$. 58 is the 64th percentile.

- b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

$x = 3$ and $y = 1. \frac{x+0.5y}{n} (100) = \frac{3+0.5(1)}{29} (100) = 12.07$. Twenty-five is the 12th percentile.

Note:

Try It

Exercise:**Problem:**

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Solution:

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$x = 15$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{15+0.5(1)}{29}(100) = 53.45$. 47 is the 53rd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$x = 15$ and $y = 2$. $\frac{x+0.5y}{n}(100) = \frac{15+0.5(2)}{29}(100) = 31.03$. 31 is the 31st percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the p th percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Note:**NOTE**

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example:**Exercise:****Problem:**

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution:

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Note:**Try It****Exercise:**

Problem:

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Solution:

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

Example:**Exercise:****Problem:**

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution:

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Note:

Try It

Exercise:**Problem:**

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Solution:

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example:**Exercise:****Problem:**

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Solution:

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Note:

Try It

Exercise:**Problem:**

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Solution:

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Example:

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- $\text{Min} = 0$
- $Q_1 = 20$
- $\text{Med} = 40$
- $Q_3 = 60$
- $\text{Max} = 300$

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes ($60 - 20 = 40$), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120.$$

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- $\text{Min} = 0$
- $Q_1 = 20$
- $Q_3 = 60$
- $\text{Max} = 120$

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1> (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/> (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/main/www/cen1990.html> (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Chapter Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile, the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3 , and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \left(\frac{k}{100} \right) (n + 1)$$

where i = the ranking or position of a data value,

k = the k th percentile,

n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x + 0.5y}{n} \right) (100)$

where x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- Find the 40th percentile.
- Find the 78th percentile.

Solution:

- The 40th percentile is 37 years.
- The 78th percentile is 70 years.

Exercise:

Problem:

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- a. Find the percentile of 37.
- b. Find the percentile of 72.

Exercise:**Problem:**

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Solution:

Jesse graduated 37th out of a class of 180 students. There are $180 - 37 = 143$ students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercise:**Problem:**

- a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
- c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise:**Problem:**

- a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
- b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:**Problem:**

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise:**Problem:**

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:**Problem:**

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise:**Problem:**

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample.

INTERPRETATION: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

Exercise:**Problem:**

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
- b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise:**Problem:**

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is \$240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more.

Use the following information to answer the next six exercises. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise:

Problem: First quartile = _____

Exercise:

Problem: Second quartile = median = 50th percentile = _____

Solution:

4

Exercise:

Problem: Third quartile = _____

Exercise:

Problem: Interquartile range (*IQR*) = _____ - _____ = _____

Solution:

$$6 - 4 = 2$$

Exercise:

Problem: 10th percentile = _____

Exercise:

Problem: 70th percentile = _____

Solution:

6

Homework

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

- Based upon this information, give two reasons why the black median age could be lower than the white median age.
- Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
- How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in [\[link\]](#). Also, include left endpoint, but not the right endpoint.

Salary (\$)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

- What percentage of the survey answered "not sure"?
- What percentage think that middle-class is from \$25,000 to \$50,000?
- Construct a histogram of the data.
 - Should all bars have the same width, based on the data? Why or why not?
 - How should the <20,000 and the 100,000+ intervals be handled? Why?
- Find the 40th and 80th percentiles
- Construct a bar graph of the data

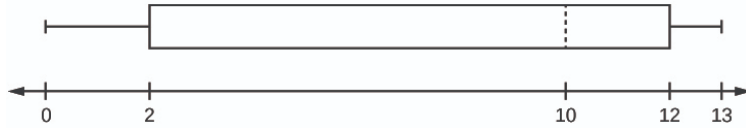
Solution:

- $1 - (0.02 + 0.09 + 0.19 + 0.26 + 0.18 + 0.17 + 0.02 + 0.01) = 0.06$
- $0.19 + 0.26 + 0.18 = 0.63$
- Check student's solution.
- 40th percentile will fall between 30,000 and 40,000
 80th percentile will fall between 50,000 and 75,000

e. Check student's solution.

Exercise:

Problem: Given the following box plot:

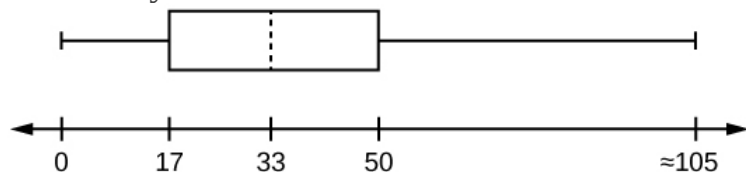


- which quarter has the smallest spread of data? What is that spread?
- which quarter has the largest spread of data? What is that spread?
- find the interquartile range (*IQR*).
- are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
- which interval has the fewest data in it? How do you know this?
 - 0–2
 - 2–4
 - 10–12
 - 12–13
 - need more information

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year.



- Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
- 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

Solution:

- a. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
- b. 62.4%

Glossary

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- **sample:** $x = \bar{x} + (\#ofSTDEV)(s)$
- **Population:** $x = \mu + (\#ofSTDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol \bar{x} is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference " $x - \text{mean}$ " is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N , the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f(x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is ***n* - 1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is ***N***, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is one. If a value appears three times in the data set or population, *f* is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in [Descriptive Statistics: Measuring the Center of the Data](#). How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter [The Central Limit Theorem](#) (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and *n* is the size of the sample.

Note:

NOTE

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example:

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year: 9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

Equation:

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s .

Data	Freq.	Deviations	<i>Deviations</i>²	(Freq.) (<i>Deviations</i>²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times 0.950625 = 2.851875$

Data	Freq.	Deviations	<i>Deviations</i> ²	(Freq.) (<i>Deviations</i> ²)
				The total is 9.7375

The sample variance, s^2 , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one ($20 - 1$):

$$s^2 = \frac{9.7375}{20-1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891, \text{ which is rounded to two decimal places, } s = 0.72.$$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Exercise:

Problem:

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
 - For a sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
 - For a population: $x = \mu + (\text{\#ofSTDEVs})(\sigma)$
 - For this example, use $x = \bar{x} + (\text{\#ofSTDEVs})(s)$ because the data is from a sample
- a. Verify the mean and standard deviation on your calculator or computer.
 - b. Find the value that is one standard deviation above the mean. Find $(\bar{x} + 1s)$.
 - c. Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.
 - d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

a. Note:

- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x} = 10.525$
- Use S_x because this is sample data (not a population): $S_x = 0.715891$

b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$

c. $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$

d. $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
 $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Note:

Try It

Exercise:

Problem: On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36;
 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Solution:

$$\mu = 30.68$$

$$s = 6.09$$

$$(\bar{x} + 2s) = 30.68 + (2)(6.09) = 42.86.$$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is -1.525 for the data value nine. **If you add the deviations, the sum is always zero.** (For [link](#), there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

Note:

NOTE

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that

the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example:

Exercise:

Problem:

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

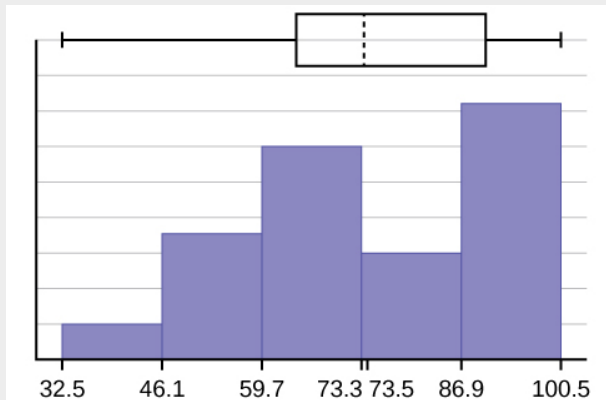
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
 - i. The sample mean
 - ii. The sample standard deviation
 - iii. The median
 - iv. The first quartile
 - v. The third quartile
 - vi. *IQR*
- c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution:

- a. See [\[link\]](#)
- b.
 - i. The sample mean = 73.5
 - ii. The sample standard deviation = 17.9
 - iii. The median = 73
 - iv. The first quartile = 61
 - v. The third quartile = 90
 - vi. $IQR = 90 - 61 = 29$

- c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5$ = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ($IQR = 29$) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Note:
Try It
Exercise:

Problem:

The following data show the different types of pet food stores in the area carry.
6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12;
12; 12; 12;
Calculate the sample mean and the sample standard deviation to one decimal
place using a TI-83+ or TI-84 calculator.

Solution:

$$\mu = 9.3$$

$$s = 2.2$$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f}$$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example:

Find the standard deviation for the data in [\[link\]](#).

Class	Frequency, f	Midpoint, m	m^2	x^2	fm^2	Standard Deviation
-------	-------------------	------------------	-------	-------	--------	-----------------------

Class	Frequency, f	Midpoint, m	m^2	x^2	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12– 14	0	13	169	7.58	0	3.5
15– 17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $x = 7.58$ and the standard deviation, $s_x = 3.5$. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since $7.58 - 3.5 - 3.5 = 0.58$.

While the formula for calculating the standard deviation is not complicated,

$s_x = \sqrt{\frac{f(m-x)^2}{n-1}}$ where s_x = sample standard deviation, x = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Note:

Try It

Find the standard deviation for the data from the previous example

Class	Frequency, f
0–2	1

Class	Frequency, f
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

First, press the **STAT** key and select **1:Edit**

```

EDIT  CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

```

Input the midpoint values into **L1** and the frequencies into **L2**

L1	L2	L3	2
1	1		
4	6		
7	10		
10	7		
13	0		
16	2		
-----	-----		

L2(7) =

Select **STAT**, **CALC**, and **1: 1-Var Stats**

```

EDIT  CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

```

Select **2nd** then **1** then , **2nd** then **2** **Enter**

```

1-Var Stats
x̄=7.576923077
Σx=197
Σx²=1799
Sx=3.500549407
σx=3.432571103
↓n=26

```

You will see displayed both a population standard deviation, σ_x , and the sample standard deviation, s_x .

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#ofSTDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z . In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \# \text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

$$\text{For John, } z = \# \text{ of STDEVs} = \frac{2.85 - 3.0}{0.7} = -0.21$$

$$\text{For Ali, } z = \# \text{ of STDEVs} = \frac{77 - 80}{10} = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of -0.21 is higher than Ali's z-score of -0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Note:

Try It

Exercise:

Problem:

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Solution:

For Angie: $z = \frac{26.2 - 27.2}{0.8} = -1.25$

For Beth: $z = \frac{27.3 - 30.1}{1.4} = -2$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.

- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

Chapter Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.

- $s = \sqrt{\frac{\sum (x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-x)^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we

would use the population mean, μ , and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} \text{ where } \begin{matrix} s_x = \text{sample standard deviation} \\ x = \text{sample mean} \end{matrix}$$

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Exercise:

Problem:

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Solution:

$$s = 34.5$$

Exercise:

Problem: Find the value that is one standard deviation below the mean.

Exercise:**Problem:**

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Solution:

$$\text{For Fredo: } z = \frac{0.158 - 0.166}{0.012} = -0.67$$

$$\text{For Karl: } z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's z-score of -0.67 is higher than Karl's z-score of -0.8 . For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise:

Problem: Use [\[link\]](#) to find the value that is three standard deviations:

- **a**above the mean
- **b**below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise:

Problem:

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53

Daily Low Temperature	Frequency
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solution:

$$\text{a. } s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$$

$$\text{b. } s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$$

$$\text{c. } s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$$

Homework

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median = 1,014 FTES
- $\sigma = 474$ FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- $n = 29$ years

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Exercise:

Problem: 75% of all years have an FTES:

- a. at or below: _____
- b. at or above: _____

Exercise:

Problem: The population standard deviation = _____

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Exercise:

Problem: What is the *IQR*? What does the *IQR* represent?

Solution:

919

Exercise:

Problem: How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Exercise:

Problem:

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

Solution:

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Exercise:

Problem:

What additional information is needed to construct a box plot for the FTES for 2005-2006 through 2010-2011 and a box plot for the FTES for 1976-1977 through 2004-2005?

Exercise:**Problem:**

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQRs* are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Exercise:**Problem:**

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Exercise:

Problem:

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing \$3,000, a guitar costing \$550, and a drum set costing \$600. The mean cost for a piano is \$4,000 with a standard deviation of \$2,500. The mean cost for a guitar is \$500 with a standard deviation of \$200. The mean cost for drums is \$700 with a standard deviation of \$100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Exercise:**Problem:**

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

- a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
- b. Who is the fastest runner with respect to his or her class? Explain why.

Exercise:**Problem:**

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in [Table 14](#).

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

Solution:

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that $23.32 + 12.95 = 36.27$ is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

Exercise:

Problem:

[\[link\]](#) gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

Bringing It Together

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4

# of movies	Frequency
4	1

- Find the sample mean \bar{x} .
- Find the approximate sample standard deviation, s .

Solution:

- 1.48
- 1.12

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

- Find the sample mean \bar{x}

- b. Find the sample standard deviation, s
- c. Construct a histogram of the data.
- d. Complete the columns of the chart.
- e. Find the first quartile.
- f. Find the median.
- g. Find the third quartile.
- h. Construct a box plot of the data.
- i. What percent of the students owned at least five pairs?
- j. Find the 40th percentile.
- k. Find the 90th percentile.
- l. Construct a line graph of the data
- m. Construct a stemplot of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

- a. Organize the data from smallest to largest value.
- b. Find the median.
- c. Find the first quartile.
- d. Find the third quartile.
- e. Construct a box plot of the data.
- f. The middle 50% of the weights are from _____ to _____.
- g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- i. Assume the population was the San Francisco 49ers. Find:
 - i. the population mean, μ .
 - ii. the population standard deviation, σ .
 - iii. the weight that is two standard deviations below the mean.
 - iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

- j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmitt Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution:

- a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
- b. 241
- c. 205.5
- d. 272.5



- f. 205.5, 272.5
- g. sample
- h. population
- i. i. 236.34
ii. 37.50
iii. 161.34
iv. 0.84 std. dev. below the mean
- j. Young

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

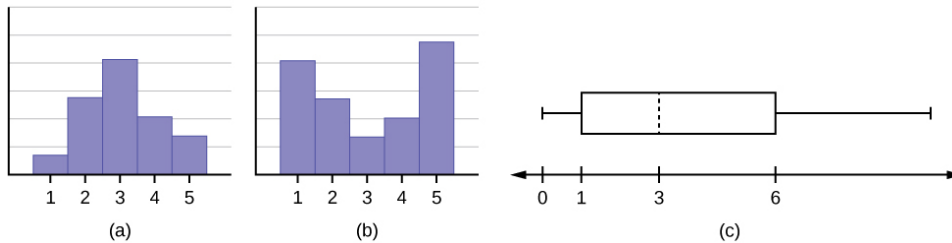
3 8-12 05-31-16 5-2

- What is the mean change score?
- What is the standard deviation for this population?
- What is the median change score?
- Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Refer to [\[link\]](#) determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



- The medians for all three graphs are the same.
- We cannot determine if any of the means for the three graphs is different.
- The standard deviation for graph b is larger than the standard deviation for graph a.
- We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

- True
- True
- True
- False

Exercise:

Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- Organize the data in a chart.

- b. Find the median, the first quartile, and the third quartile.
- c. Find the 65th percentile.
- d. Find the 10th percentile.
- e. Construct a box plot of the data.
- f. The middle 50% of the conferences last from _____ days to _____ days.
- g. Calculate the sample mean of days of engineering conferences.
- h. Calculate the sample standard deviation of days of engineering conferences.
- i. Find the mode.
- j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

- a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
- b. Construct a histogram of the data.
- c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- d. Calculate the sample mean.
- e. Calculate the sample standard deviation.
- f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Solution:

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.

c. mode

d. 8628.74

e. 6943.88

f. -0.09

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11

x	Frequency
5	9
6	4

Exercise:

Problem: The 80th percentile is _____

- a. 5
- b. 80
- c. 3
- d. 4

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately _____

- a. 0.7
- b. 4.8
- c. -2.8
- d. Cannot be determined

Solution:

a

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the [\[link\]](#).

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

- Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
- If a data value is identified as an outlier, what should be done about it?
- Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
- Do parts a and c of this problem give the same answer?
- Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
- Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

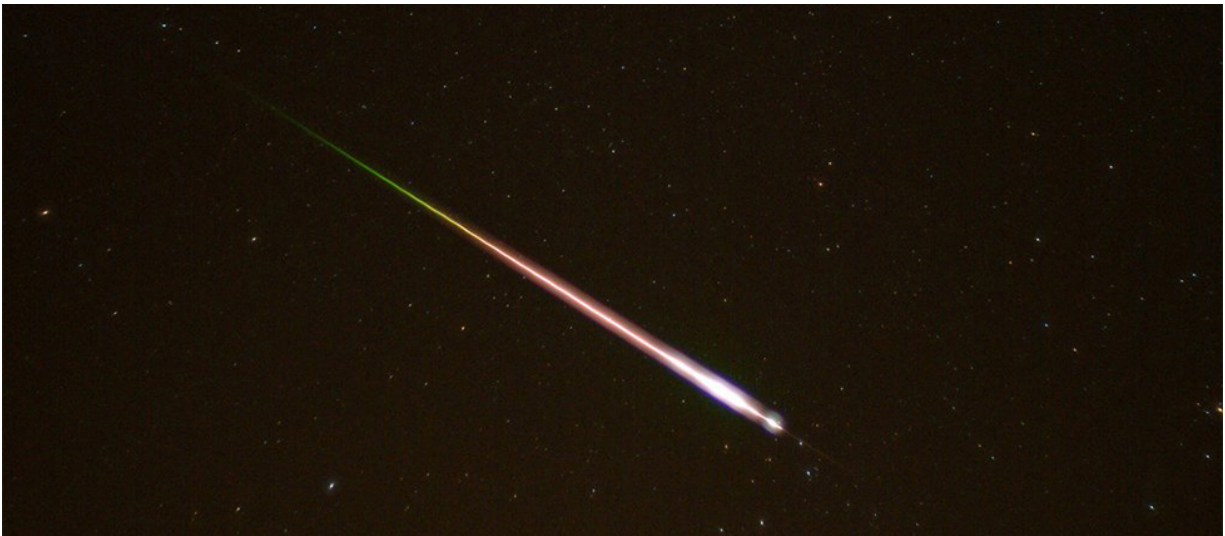
Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and \bar{x} is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Introduction

class="introduction"

Meteor
showers are
rare, but the
probability of
them occurring
can be
calculated.
(credit:
Navicore/flickr
)



Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.

- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Note:**Collaborative Exercise**

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A

and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers $\{1, 2, 3, 4, 5, 6\}$ on its faces. Let event E = rolling a number that is at least five. There are two outcomes $\{5, 6\}$. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces.

Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

"OR" Event:

An outcome is in the event $A \text{ OR } B$ if the outcome is in A or is in B or is in both A and B . For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. $A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

"AND" Event:

An outcome is in the event $A \text{ AND } B$ if the outcome is in both A and B at the same time. For example, let A and B be $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7, 8\}$, respectively. Then $A \text{ AND } B = \{4, 5\}$.

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A . Notice that $P(A) + P(A') = 1$. For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = \{1, 2, 3, 4\}$. Then, $A' = \{5, 6\}$. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

The **conditional probability** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space.** We calculate the probability of A from the reduced sample space B . The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let $A = \text{face is 2 or 3}$ and $B = \text{face is even (2, 4, 6)}$. To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{the number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example:

Exercise:

Problem:

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____

Let event A = the even numbers and event B = numbers greater than 13.

b. $A =$ _____, $B =$ _____

c. $P(A) =$ _____, $P(B) =$ _____

d. $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____

e. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____

f. $A' =$ _____, $P(A') =$ _____

g. $P(A) + P(A') =$ _____

h. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Solution:

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$

b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$

c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$

d. $A \text{ AND } B = \{14, 16, 18\}$, $A \text{ OR } B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$

e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$

f. $A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(A') = \frac{10}{19}$

g. $P(A) + P(A') = 1 \left(\frac{9}{19} + \frac{10}{19} = 1 \right)$

h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{3}{9}$, No

Note:

Try It

Exercise:**Problem:**

The sample space S is all the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ _____

Let event A = the sum is even and event B = the first number is prime.

b. $A =$ _____, $B =$ _____

c. $P(A) =$ _____, $P(B) =$ _____

d. $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____

e. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____

f. $B' =$ _____, $P(B') =$ _____

- g. $P(A) + P(A') =$ _____
 h. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Solution:

- a. $S = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
 b. $A = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$
 $B = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
 c. $P(A) = \frac{1}{2}, P(B) = \frac{2}{3}$
 d. $A \text{ AND } B = \{(2,2), (2,4), (3,1), (3,3)\}$
 $A \text{ OR } B = \{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
 e. $P(A \text{ AND } B) = \frac{1}{3}, P(A \text{ OR } B) = \frac{5}{6}$
 f. $B' = \{(1,1), (1,2), (1,3), (1,4)\}, P(B') = \frac{1}{3}$
 g. $P(B) + P(B') = 1$
 h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{1}{2}, P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{2}{3}, \text{ No.}$

Example:

Exercise:

Problem:

A fair, six-sided die is rolled. Describe the sample space S , identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

- a. Event T = the outcome is two.
 b. Event A = the outcome is an even number.
 c. Event B = the outcome is less than four.

- d. The complement of A .
- e. A GIVEN B
- f. B GIVEN A
- g. A AND B
- h. A OR B
- i. A OR B'
- j. Event N = the outcome is a prime number.
- k. Event I = the outcome is seven.

Solution:

- a. $T = \{2\}, P(T) = \frac{1}{6}$
- b. $A = \{2, 4, 6\}, P(A) = \frac{1}{2}$
- c. $B = \{1, 2, 3\}, P(B) = \frac{1}{2}$
- d. $A' = \{1, 3, 5\}, P(A') = \frac{1}{2}$
- e. $A|B = \{2\}, P(A|B) = \frac{1}{3}$
- f. $B|A = \{2\}, P(B|A) = \frac{1}{3}$
- g. $A \text{ AND } B = \{2\}, P(A \text{ AND } B) = \frac{1}{6}$
- h. $A \text{ OR } B = \{1, 2, 3, 4, 6\}, P(A \text{ OR } B) = \frac{5}{6}$
- i. $A \text{ OR } B' = \{2, 4, 5, 6\}, P(A \text{ OR } B') = \frac{2}{3}$
- j. $N = \{2, 3, 5\}, P(N) = \frac{1}{2}$
- k. A six-sided die does not have seven dots. $P(7) = 0$.

Example:

[\[link\]](#) describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Exercise:

Problem:

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

- $P(M)$
- $P(F)$
- $P(R)$
- $P(L)$
- $P(M \text{ AND } R)$
- $P(F \text{ AND } L)$
- $P(M \text{ OR } F)$
- $P(M \text{ OR } R)$
- $P(F \text{ OR } L)$
- $P(M')$
- $P(R|M)$
- $P(F|L)$
- $P(L|F)$

Solution:

- $P(M) = 0.52$
- $P(F) = 0.48$
- $P(R) = 0.87$
- $P(L) = 0.13$
- $P(M \text{ AND } R) = 0.43$
- $P(F \text{ AND } L) = 0.04$

- g. $P(M \text{ OR } F) = 1$
- h. $P(M \text{ OR } R) = 0.96$
- i. $P(F \text{ OR } L) = 0.57$
- j. $P(M') = 0.48$
- k. $P(R|M) = 0.8269$ (rounded to four decimal places)
- l. $P(F|L) = 0.3077$ (rounded to four decimal places)
- m. $P(L|F) = 0.0833$

References

“Countries List by Continent.” Worldatlas, 2013. Available online at <http://www.worldatlas.com/cntycont.htm> (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

Exercise:

Problem:

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
 - Let M be the event that a student is male.
 - Let S be the event that a student has short hair.
 - Let L be the event that a student has long hair.
-
- a. The probability that a student does not have long hair.
 - b. The probability that a student is male or has short hair.
 - c. The probability that a student is a female and has long hair.
 - d. The probability that a student is male, given that the student has long hair.
 - e. The probability that a student has long hair, given that the student is male.
 - f. Of all the female students, the probability that a student has short hair.
 - g. Of all students with long hair, the probability that a student is female.
 - h. The probability that a student is female or has long hair.
 - i. The probability that a randomly selected student is a male student with short hair.
 - j. The probability that a student is female.

Solution:

- a. $P(L') = P(S)$
- b. $P(M \text{ OR } S)$
- c. $P(F \text{ AND } L)$
- d. $P(M|L)$

- e. $P(L|M)$
- f. $P(S|F)$
- g. $P(F|L)$
- h. $P(F \text{ OR } L)$
- i. $P(M \text{ AND } S)$
- j. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let H = the event of getting a hat.

Let N = the event of getting a noisemaker.

Let F = the event of getting a finger trap.

Let C = the event of getting a bag of confetti.

Exercise:

Problem: Find $P(H)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(C)$.

Solution:

$$P(C) = \frac{5}{42} = 0.12$$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let B = the event of getting a blue jelly bean

Let G = the event of getting a green jelly bean.

Let O = the event of getting an orange jelly bean.

Let P = the event of getting a purple jelly bean.

Let R = the event of getting a red jelly bean.

Let Y = the event of getting a yellow jelly bean.

Exercise:

Problem: Find $P(B)$.

Exercise:

Problem: Find $P(G)$.

Solution:

$$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$$

Exercise:

Problem: Find $P(P)$.

Exercise:

Problem: Find $P(R)$.

Solution:

$$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$$

Exercise:

Problem: Find $P(Y)$.

Exercise:

Problem: Find $P(O)$.

Solution:

$$P(O) = \frac{150 - 22 - 38 - 20 - 28 - 26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let A = the event that a country is in Asia.

Let E = the event that a country is in Europe.

Let F = the event that a country is in Africa.

Let N = the event that a country is in North America.

Let O = the event that a country is in Oceania.

Let S = the event that a country is in South America.

Exercise:

Problem: Find $P(A)$.

Exercise:

Problem: Find $P(E)$.

Solution:

$$P(E) = \frac{47}{194} = 0.24$$

Exercise:

Problem: Find $P(F)$.

Exercise:**Problem:** Find $P(N)$.**Solution:**

$$P(N) = \frac{23}{194} = 0.12$$

Exercise:**Problem:** Find $P(O)$.**Exercise:****Problem:** Find $P(S)$.**Solution:**

$$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$$

Exercise:**Problem:**

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise:**Problem:**

What is the probability of drawing a club in a standard deck of 52 cards?

Solution:

$$\frac{13}{52} = \frac{1}{4} = 0.25$$

Exercise:

Problem:

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise:**Problem:**

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Solution:

$$\frac{3}{6} = \frac{1}{2} = 0.5$$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



Let B = the event of landing on blue.
Let R = the event of landing on red.

Let G = the event of landing on green.

Let Y = the event of landing on yellow.

Exercise:

Problem: If you land on Y , you get the biggest prize. Find $P(Y)$.

Exercise:

Problem: If you land on red, you don't get a prize. What is $P(R)$?

Solution:

$$P(R) = \frac{4}{8} = 0.5$$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let I = the event that a player is an infielder.

Let O = the event that a player is an outfielder.

Let H = the event that a player is a great hitter.

Let N = the event that a player is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is not an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Solution:

$$P(O \text{ OR } H)$$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Solution:

$$P(H|I)$$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Solution:

$$P(N|O)$$

Exercise:

Problem:

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Solution:

$$P(I \text{ OR } N)$$

Exercise:**Problem:**

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise:**Problem:**

Write the symbols for the probability that a player is an infielder.

Solution:

$$P(I)$$

Exercise:

Problem: What is the word for the set of all possible outcomes?

Exercise:

Problem: What is conditional probability?

Solution:

The likelihood that an event will occur given that another event has already occurred.

Exercise:

Problem:

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book
Let F = event that book is fiction
Let N = event that book is nonfiction
What is the sample space?

Exercise:**Problem:**

What is the sum of the probabilities of an event and its complement?

Solution:

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

Exercise:

Problem: What does $P(E|M)$ mean in words?

Exercise:

Problem: What does $P(E \text{ OR } M)$ mean in words?

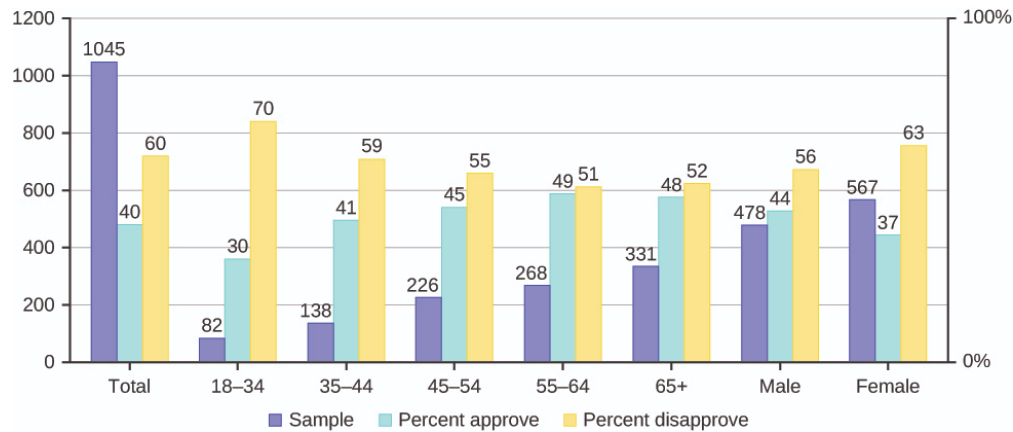
Solution:

the probability of landing on an even number or a multiple of three

Homework

Exercise:

Problem:



The graph in [\[link\]](#) displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

- Define three events in the graph.
- Describe in words what the entry 40 means.
- Describe in words the complement of the entry in question 2.
- Describe in words what the entry 30 means.
- Out of the males and females, what percent are males?
- Out of the females, what percent disapprove of Mayor Ford?
- Out of all the age groups, what percent approve of Mayor Ford?
- Find $P(\text{Approve}|\text{Male})$.
- Out of the age groups, what percent are more than 44 years old?
- Find $P(\text{Approve}|\text{Age} < 35)$.

Exercise:

Problem:

Explain what is wrong with the following statements. Use complete sentences.

- a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
 - b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.
-

Solution:

- a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
- b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S . An event is an arbitrary subset in S . It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A , B , C , and so on.

Experiment

a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1$
- If A and B are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(B)$.
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event $A \text{ AND } B$ if the outcome is in both $A \text{ AND } B$ at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A .

The Conditional Probability of A GIVEN B

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event $A \text{ OR } B$ if the outcome is in A or is in B or is in both A and B .

Introduction

class="introduction"

Linear
regression
and
correlation
can help
you
determine
if an auto
mechanic's
salary is
related to
his work
experience
. (credit:
Joshua
Rothhaas)



Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y . The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

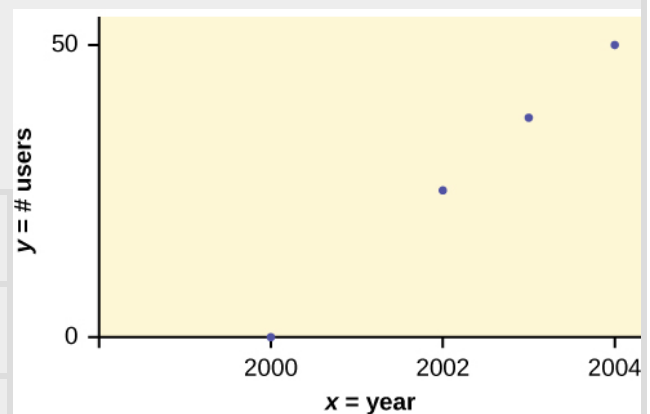
Example:

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0

Scatter plot showing the number of m-commerce users (in millions) by year.



x (year)	y (# of users)
2004	47.0

Note: To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Note:

Try It

Exercise:

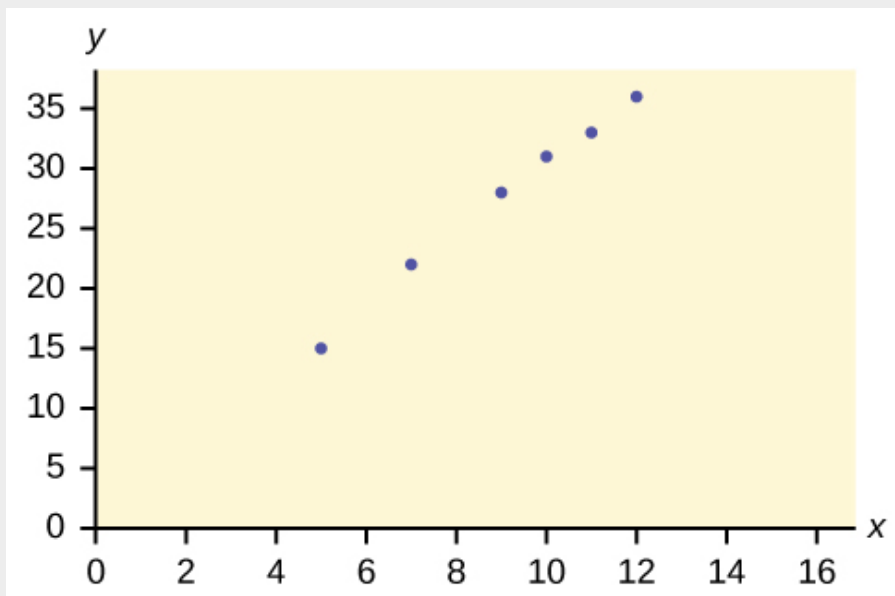
Problem:

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Solution:



Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

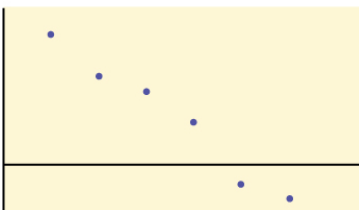
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



(a) Positive linear pattern (strong)



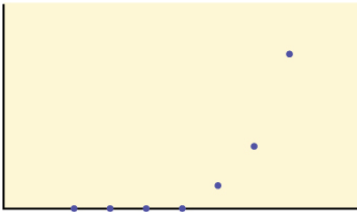
(b) Linear pattern w/ one deviation



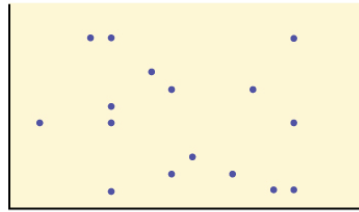
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x .

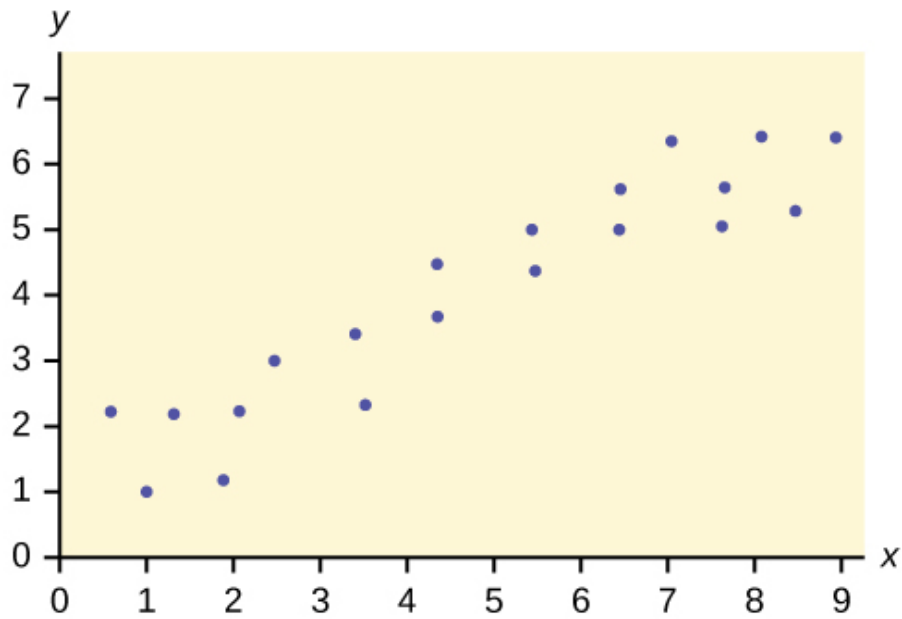
Chapter Review

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



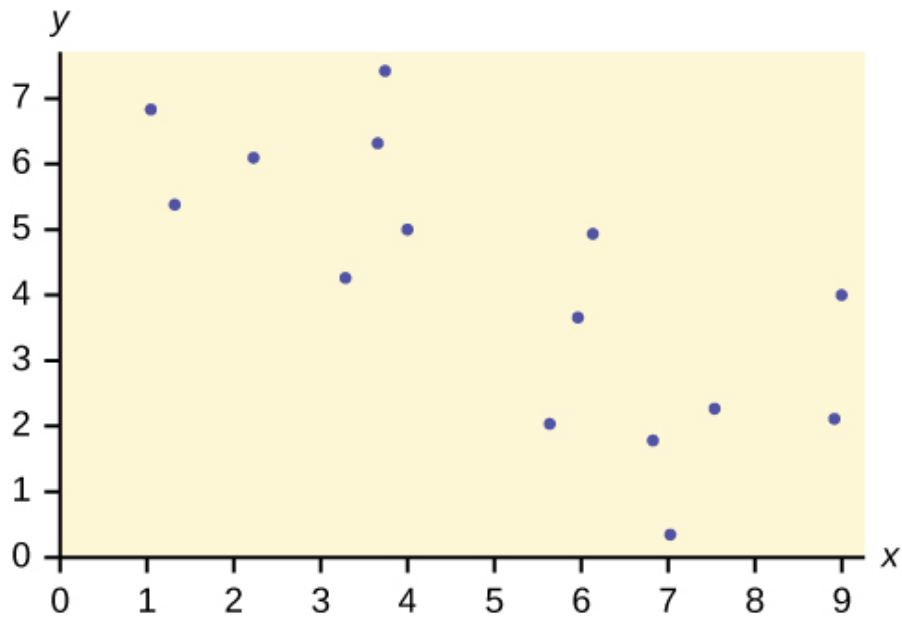
Solution:

The data appear to be linear with a strong, positive correlation.

Exercise:

Problem:

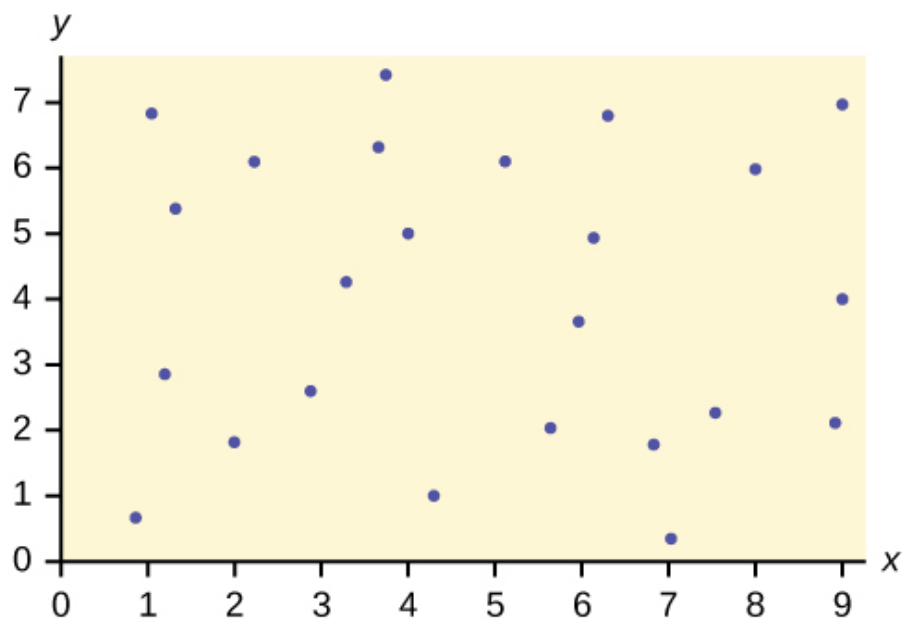
Does the scatter plot appear linear? Strong or weak? Positive or negative?



Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



Solution:

The data appear to have no correlation.

Homework

Exercise:

Problem:

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. [\[link\]](#) shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Solution:

Check student's solution.

Exercise:**Problem:**

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Exercise:**Problem:**

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
--------	----------------------------------	----------------

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Solution:

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Exercise:

Problem:

If the level of significance is 0.05 and the p -value is 0.06, what conclusion can you draw?

Exercise:**Problem:**

If there are 15 data points in a set of data, what is the number of degree of freedom?

Solution:

13

Testing the Significance of the Correlation Coefficient

The correlation coefficient, r , tells us about the strength and direction of the linear relationship between x and y . However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n , together.

We perform a hypothesis test of the "**significance of the correlation coefficient**" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r , the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r , is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ , the Greek letter "rho."
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n .

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

- What the conclusion means: There is a significant linear relationship between x and y . We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y . Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

Note:

Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- Null Hypothesis: $H_0: \rho = 0$
- Alternate Hypothesis: $H_a: \rho \neq 0$

WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis H_0 :** The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis H_a :** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:

There are two methods of making the decision. The two methods are equivalent and give the same result.

- **Method 1: Using the p -value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

Note:

Note

Using the p -value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p -value to make a decision

Note:

To calculate the p -value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for β or ρ , highlight " $\neq 0$ "

The output screen shows the p -value on the line that reads "p =".

(Most computer statistical software can calculate the p -value.)

If the p -value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p -value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p -value. The following describes the calculations to compute the test statistics and the p -value:
- The p -value is calculated using a t -distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t , is shown in the computer or calculator output along with the p -value. The test statistic t has the same sign as the correlation coefficient r .
- The p -value is the combined area in both tails.

An alternative way to calculate the p -value (**p**) given by LinRegTTest is the command `2*tcdf(abs(t),10^99, n-2)` in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p -value method

- Consider the [third exam/final exam example](#).
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$$H_0: \rho = 0$$

$$H_a: \rho \neq 0$$

$$\alpha = 0.05$$

- The p -value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p -value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

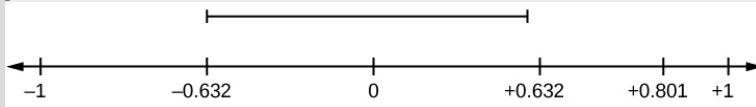
Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The [95% Critical Values of the Sample Correlation Coefficient Table](#) can be used to give you a good idea of whether the computed value of r is **significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example:

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r < \text{negative critical value}$ or $r > \text{positive critical value}$, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.



r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Note:

Try It

Exercise:

Problem:

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576 . Can the line be used for prediction? Why or why not?

Solution:

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r > \text{the positive critical value}$.

Example:

Suppose you computed $r = -0.624$ with 14 data points. $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532 . Since $-0.624 < -0.532$, r is

significant and the line can be used for prediction



$r = -0.624 < -0.532$. Therefore, r is significant.

Note:

Try It

Exercise:

Problem:

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Solution:

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example:

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811 . Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.



$-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

Note:

Try It

Exercise:**Problem:**

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is $= 0.707$. Can the line be used for prediction? Why or why not?

Solution:

Yes, the line can be used for prediction, because $r < \text{the negative critical value}$.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the [third exam/final exam example](#). The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- $H_0: \rho = 0$
- $H_a: \rho \neq 0$
- $\alpha = 0.05$
- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are -0.602 and $+0.602$
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example:

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

- a. $r = -0.567$ and the sample size, n , is 19. The $df = n - 2 = 17$. The critical value is -0.456 . $-0.567 < -0.456$ so r is significant.
- b. $r = 0.708$ and the sample size, n , is nine. The $df = n - 2 = 7$. The critical value is 0.666 . $0.708 > 0.666$ so r is significant.
- c. $r = 0.134$ and the sample size, n , is 14. The $df = 14 - 2 = 12$. The critical value is 0.532 . 0.134 is between -0.532 and 0.532 so r is not significant.
- d. $r = 0$ and the sample size, n , is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

Note:

Try It

Exercise:

Problem:

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Solution:

No, the line cannot be used for prediction no matter what the sample size is.

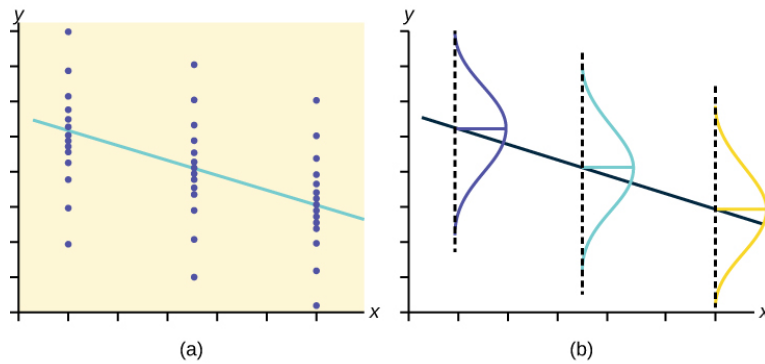
Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x . In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x . In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.



The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Chapter Review

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope **b** and intercept **a** of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ , use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation

coefficient. To test the null hypothesis $H_0: \rho = \text{hypothesized value}$, use a linear regression t-test. The most common null hypothesis is $H_0: \rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx$$

where

a = y-intercept

b = slope

Standard deviation of the residuals:

$$s = \sqrt{\frac{SEE}{n-2}}.$$

where

SSE = sum of squared errors

n = the number of data points

Exercise:

Problem:

When testing the significance of the correlation coefficient, what is the null hypothesis?

Exercise:

Problem:

When testing the significance of the correlation coefficient, what is the alternative hypothesis?

Solution:

$$H_a: \rho \neq 0$$

Exercise:**Problem:**

If the level of significance is 0.05 and the p -value is 0.04, what conclusion can you draw?

Review Exercises (Ch 3-13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3, cover material learned in chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks.

1. The “average increase” for all NASDAQ stocks is the:

- a. population
- b. statistic
- c. parameter
- d. sample
- e. variable

2. All of the NASDAQ stocks are the:

- a. population
- b. statistics
- c. parameter
- d. sample
- e. variable

3. Nine percent is the:

- a. population
- b. statistics
- c. parameter
- d. sample
- e. variable

4. The 100 NASDAQ stocks in the survey are the:

- a. population
- b. statistic
- c. parameter
- d. sample
- e. variable

5. The percent increase for one stock in the survey is the:

- a. population
- b. statistic
- c. parameter
- d. sample
- e. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Weight Gain	Frequency
-2	3
-1	5
0	2
1	4
4	13
6	2

Weight Gain	Frequency
11	1

7. Calculate the following values:

- the average weight gain for the two weeks
- the standard deviation
- the first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises: A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than \$2,000 per month. Of those respondents who charge more than \$2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than \$2,000 AND use a credit card that gives them a mile of air travel for every dollar they charge?

- $(0.30)(0.35)$
- $(0.80)(0.35)$
- $(0.80)(0.30)$
- (0.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent AND charging more than \$2,000 per month independent events?

- Yes
- No, and they are not mutually exclusive either.
- No, but they are mutually exclusive.
- Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

- a. all employed adult women
- b. all the members of a local business and professional women's club
- c. the 100 women who received the questionnaire
- d. all employed women with children

Use the following information to answer the next two exercises: The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

Gender/Ethnic group	White	Asian	Hispanic	Black	American Indian
Male	400	468	115	35	16
Female	440	132	140	40	14

12. Find the probability that a student is Asian or Male.

13. Find the probability that a student is Black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs.
- Median = 4.5 lbs.

- Mode = 4 lbs.
- Standard deviation = 3.8 lbs.
- First quartile = 2 lbs.
- Third quartile = 8.5 lbs.

The correct statement is:

- One fourth of the members lost exactly two pounds.
- The middle fifty percent of the members lost from two to 8.5 lbs.
- Most people lost 3.5 to 4.5 lbs.
- All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

- All values of the data appear with the same frequency.
- The mean of the data is also zero.
- All of the data have the same value.
- There are no data to begin with.

16. The statement that describe the illustration is:



- the mean is equal to the median.
- There is no first quartile.
- The lowest data value is the median.
- The median equals $\frac{Q_1 + Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss (deafness) is approximately 2 per 1000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby

nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A “friend” offers you the following “deal.” For a \$10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth \$6.
- Eighty of the coupons are for a free gift worth \$8.
- Six of the coupons are for a free gift worth \$12.
- Four of the coupons are for a free gift worth \$40.

Based upon the financial gain or loss over the long run, should you play the game?

- a. Yes, I expect to come out ahead in money.
- b. No, I expect to come out behind in money.
- c. It doesn't matter. I expect to break even.

Use the following information to answer the next four exercises: Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X .

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

- a. qualitative
- b. quantitative continuous
- c. quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next seven exercises: A recent study of mothers of junior high school children in Santa Clara County reported that 76% of the mothers are employed in paid positions. Of those mothers who are employed, 64% work full-time (over 35 hours per week), and 36% work part-time. However, out of all of the mothers in the population, 49% work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

- a. Find the percent of all mothers in the population that are NOT employed.
- b. Find the percent of mothers in the population that are employed part-time.

26. The “type of employment” is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time:

- a. mutually exclusive events? Why or why not?
- b. independent events? Why or why not?

Use the following additional information to answer the next two exercises: We randomly pick ten mothers from the above population. We are interested in the

number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X .

31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

- Define X .
- What are the values that the random variable may take on?
- State the distribution for X .
- Find the probability that from ten to 14 (inclusive) questions are posted to the listserv on a randomly picked day.

33. A person invests \$1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year (i.e. his stock will be worthless) is 35%. The probability that the person's stock will still have a value of \$1,000 after one year (i.e. no profit and no loss) is 60%. The probability that the person's stock will increase in value by \$10,000 after one year (i.e. will be worth \$11,000) is 5%. Find the expected profit after one year.

34. Rachel's piano cost \$3,000. The average cost for a piano is \$4,000 with a standard deviation of \$2,500. Becca's guitar cost \$550. The average cost for a guitar is \$500 with a standard deviation of \$200. Matt's drums cost \$600. The average cost for drums is \$700 with a standard deviation of \$100. Whose cost was lowest when compared to his or her own instrument?

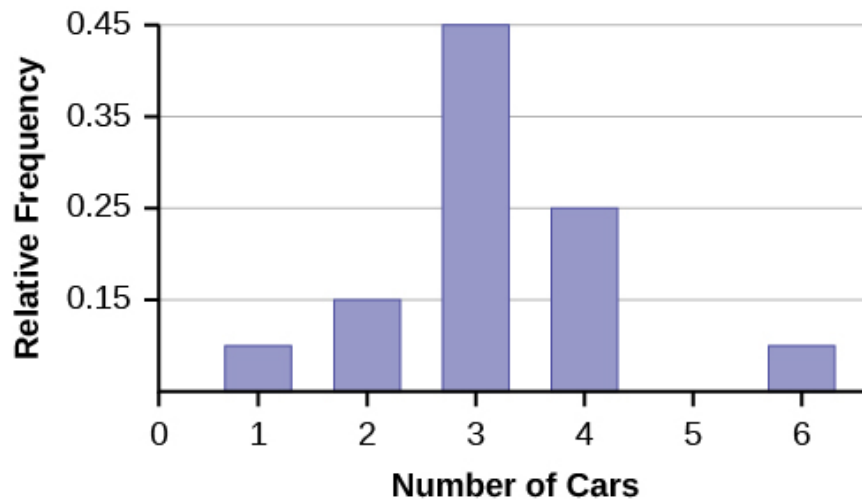


35. Explain why each statement is either true or false given the box plot in [\[link\]](#).

- Twenty-five percent of the data are at most five.
- There is the same amount of data from 4–5 as there is from 5–7.
- There are no data values of three.

d. Fifty percent of the data are four.

Using the following information to answer the next two exercises: 64 faculty members were asked the number of cars they owned (including spouse and children's cars). The results are given in the following graph:



36. Find the approximate number of responses that were three.

37. Find the first, second and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises: [\[link\]](#) shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	Blond	Brown	Black
Ponytail	3	2	5
Plain	2	2	1

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.
39. Find the probability that the girl wears her hair plain OR has brown hair.
40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

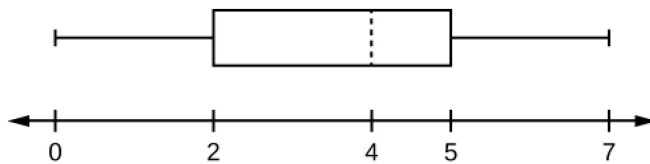
Use the following information to answer the next two exercises: $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

- a. $f(x) = \frac{1}{10}, 3 \leq x \leq 13$
- b. There is no mode
- c. The median is less than the mean.
- d. $P(x > 10) = P(x \leq 6)$

42. Calculate:

- a. the mean.
- b. the median.
- c. the 65th percentile.



43. Which of the following is true for the box plot in [\[link\]](#)?

- a. Twenty-five percent of the data are at most five.
- b. There is about the same amount of data from 4–5 as there is from 5–7.
- c. There are no data values of three.
- d. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

- a. G and H are mutually exclusive events.
- b. $P(G) = P(H)$
- c. Knowing that H has occurred will affect the chance that G will happen.
- d. G and H are independent events.

45. If $P(J) = 0.3$, $P(K) = 0.63$, and J and K are independent events, then explain which are correct and which are incorrect.

- a. $P(J \text{ AND } K) = 0$
- b. $P(J \text{ OR } K) = 0.9$
- c. $P(J \text{ OR } K) = 0.72$
- d. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X ?

- a. $P(5)$
- b. $B(500, 5)$
- c. $\text{Exp}\left(\frac{1}{5}\right)$
- d. $N\left(5, \frac{(0.01)(0.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

47. $X \sim$ _____

- a. $U(0, 4)$
- b. $U(10, 20)$

- c. $Exp(2)$
- d. $N(2, 1)$

48. The average wait time is:

- a. 1 hour.
- b. 2 hours.
- c. 2.5 hours.
- d. 4 hours.

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is:

- a. $\frac{1}{4}$
- b. $\frac{1}{2}$
- c. $\frac{3}{4}$
- d. $\frac{3}{8}$

50. Given: $X \sim Exp\left(\frac{1}{3}\right)$

- a. Find $P(x > 1)$.
- b. Calculate the minimum value for the upper quartile.
- c. Find $P\left(x = \frac{1}{3}\right)$

51.

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate
- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

- a. 4 years

- b. 4.5 years
- c. 5 years
- d. 5.5 years

52. Which of the following distributions is described by the following example?
Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

- a. binomial
- b. uniform
- c. exponential
- d. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

- a. 0.5
- b. $\frac{3}{4}$
- c. 0.43
- d. 0.63

54. Which distribution accurately describes the following situation?

The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20%. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

- a. $B(14, 0.20)$
- b. $P(2.8)$
- c. $N(2.8, 2.24)$
- d. $Exp\left(\frac{1}{0.20}\right)$

55. A 2008 report on technology use states that approximately 20% of U.S. households have never sent an e-mail. Suppose that we select a random sample of

fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email

- a. $B(14, 0.20)$
- b. $P(2.8)$
- c. $N(2.8, 2.24)$
- d. $Exp\left(\frac{1}{0.20}\right)$

Chapter 8

Use the following information to answer the next three exercises: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

- a. number of people who lost weight on the special weight loss diet.
- b. the number of people who were on the diet.
- c. the mean amount of weight lost by 15 people on the special weight loss diet.
- d. the total amount of weight lost by 15 people on the special weight loss diet.

57. Find the probability asked for in [Question 56](#).

58. Find the 90th percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

- a. cannot be determined from the information given
- b. $\frac{1}{6}$

- c. $\frac{1}{2}$
- d. $\frac{1}{3}$

60. The 20th percentile occurs after how many hours?

- a. 0.20
- b. 0.60
- c. 0.50
- d. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

- a. $U(0,3)$
- b. $Exp(13)$
- c. $N(60, 5.477)$
- d. $N(1.5, 0.01875)$

62. Using the information in [Question 61](#), find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

- a. eight days
- b. three days
- c. 14 days
- d. six days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

- a. 0.5174
- b. 0.0174
- c. 0.7500
- d. 0.2143

Use the following information to answer the next five exercises: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

- a. 80%
- b. 20%
- c. 30%
- d. cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

- a. 50%
- b. 35%
- c. 27.5%
- d. 75%

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

- a. yes
- b. no
- c. cannot be determined

68. The interest rate charged on the financial aid is _____ data.

- a. quantitative discrete
- b. quantitative continuous
- c. qualitative discrete
- d. qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = \$250
- 2nd quartile = \$700
- 3rd quartile = \$1200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive \$250 or more?

- a. 50
- b. 250
- c. 150
- d. cannot be determined

Use the following information to answer the next two exercises: $P(A) = 0.2$, $P(B) = 0.3$; A and B are independent events.

70. $P(A \text{ AND } B) = \underline{\hspace{2cm}}$

- a. 0.5
- b. 0.6
- c. 0
- d. 0.06

71. $P(A \text{ OR } B) = \underline{\hspace{2cm}}$

- a. 0.56
- b. 0.5
- c. 0.44
- d. 1

72. If H and D are mutually exclusive events, $P(H) = 0.25$, $P(D) = 0.15$, then $P(H|D)$.

- a. 1
- b. 0
- c. 0.40
- d. 0.0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt's height is two standard deviations below the mean for 14 year old boys' height. Rebecca's height is 0.10 standard deviations above the mean for 14 year old girls' height. Interpret this.

- a. Matt is 2.1 inches shorter than Rebecca.
- b. Rebecca is very tall compared to other 14 year old girls.
- c. Rebecca is taller than Matt.
- d. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see [\[link\]](#)).

Use the following information to answer the next three exercises: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	
2	0.2	
4	0.4	

x	Relative Frequency	Cumulative Relative Frequency
5	0.1	

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b) and the percent of the estimates that fall at or below four (c).

77. Calculate the median, M , the first quartile, Q_1 , the third quartile, Q_3 . Then construct a box plot of the data.

78. The middle 50% of the data are between _____ and _____.

Use the following information to answer the next three exercises: Seventy 5th and 6th graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried shrimp
5th grader	15	6	9	0
6th grader	15	7	10	8

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

- a. $\frac{32}{70}$
- b. $\frac{8}{32}$
- c. $\frac{8}{8}$
- d. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

- a. $\frac{30}{70}$
- b. $\frac{30}{40}$
- c. $\frac{40}{70}$
- d. 1

81. Find the probability a child is in the 5th grade given that the child prefers spaghetti.

- a. $\frac{9}{19}$
- b. $\frac{9}{70}$
- c. $\frac{9}{30}$
- d. $\frac{19}{70}$

82. A sample of convenience is a random sample.

- a. true
- b. false

83. A statistic is a number that is a property of the population.

- a. true
- b. false

84. You should always throw out any data that are outliers.

- a. true
- b. false

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

- a. Define the random variable X .

- b. State the distribution for X .
- c. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, 5. Assume that the underlying distribution is normal. Calculate a 95% confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

- a. mean = median \neq mode
- b. mean > median > mode
- c. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises: In a survey at Kirkwood Ski Resort the following information was recorded:

	0–10	11–20	21–40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Suppose that one person from [\[link\]](#) was randomly selected.

- 88.** Find the probability that the person was a skier or was age 11–20.
- 89.** Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

- a. Sport and age are independent events.
- b. Ski and age 11–20 are mutually exclusive events.
- c. $P(\text{Ski AND age 21–40}) < P(\text{Ski}|\text{age 21–40})$
- d. $P(\text{Snowboard OR age 0–10}) < P(\text{Snowboard}|\text{age 0–10})$

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

92. The distribution for X is uniform. What can we say for certain about the distribution for X when $n = 1$?

- a. The distribution for X is still uniform with the same mean and standard deviation as the distribution for X .
- b. The distribution for X is normal with the different mean and a different standard deviation as the distribution for X .
- c. The distribution for X is normal with the same mean but a larger standard deviation than the distribution for X .
- d. The distribution for X is normal with the same mean but a smaller standard deviation than the distribution for X .

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

- a. distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X .
- b. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for X .
- c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X .
- d. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for X .

Use the following information to answer the next three exercises: A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for \bar{X} , the distribution for the average lengths of baby carrots in 16 five-pound bags. $\bar{X} \sim$ _____

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that x is between two and 2.25 inches.

Use the following information to answer the next three exercises: At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90th percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises: Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars **at most** how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car **less than** 2.5 years.

102. If we are to pick individuals ten at a time, find the distribution for the **mean** car length ownership.

103. If we are to pick ten individuals, find the probability that the **sum** of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

- a. Uniform
- b. Exponential
- c. Normal
- d. Student t

105. Compare the standard normal distribution to the Student's t -distribution, centered at zero. Explain which of the following are true and which are false.

- a. As the number surveyed increases, the area to the left of -1 for the Student's t -distribution approaches the area for the standard normal distribution.
- b. As the degrees of freedom decrease, the graph of the Student's t -distribution looks more like the graph of the standard normal distribution.
- c. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises: We are interested in the checking account balance of twenty-year-old college students. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of \$640 and a sample standard deviation of \$150. Let X = checking account balance of an individual twenty year old college student.

106. Explain why we cannot determine the distribution of X .

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of twenty-year-old college students, what distribution would you use?

108. Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of twenty-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

- a. Define the random variable X .

- b. State the distribution for X .
- c. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises: The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- it costs \$10 to enter.
- a fair coin is tossed four times.
- if you do not get four heads or four tails, you lose your \$10.
- if you get four heads or four tails, you get back your \$10, plus \$30 more.

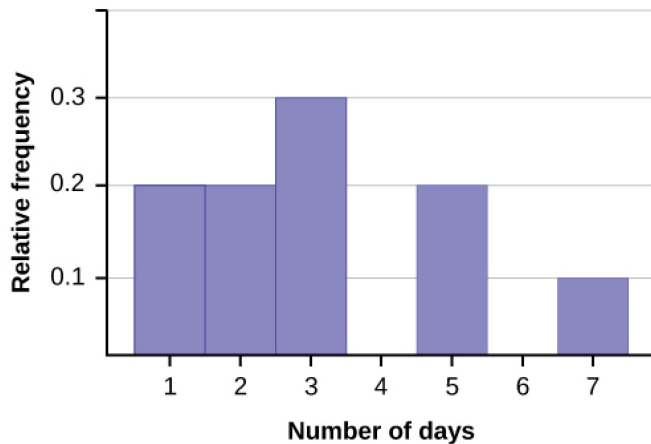
Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises: A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:



116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

- State an appropriate null hypothesis.
- State an appropriate alternative hypothesis.
- Define the random variable, P' .
- Calculate the test statistic.
- Calculate the p -value.
- At the 5% level of decision, what is your decision about the null hypothesis?
- What is the Type I error?
- What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. Fifteen Californians wear jeans

to work and six non-Californians wear jeans to work.

Let C = Californian professional; NC = non-Californian professional

- a. State appropriate null and alternate hypotheses.
- b. Define the random variable.
- c. Calculate the test statistic and p -value.
- d. At the 5% significance level, what is your decision?
- e. What is the Type I error?
- f. What is the Type II error?

Use the following information to answer the next two exercises: A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

121. This is a test of (pick the best answer):

- a. large samples, independent means
- b. small samples, independent means
- c. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises: A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12–19. Six percent of the girls surveyed said they have been pregnant. We are interested in the true proportion of U.S. girls, age 12–19, who have been pregnant.

123. Find the 95% confidence interval for the true proportion of U.S. girls, age 12–19, who have been pregnant.

124. The report also stated that the results of the survey are accurate to within $\pm 3.7\%$ at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

125. Given: $X \sim \text{Exp}\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises: The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is \$72.

126. Find the probability that one customer spends less than \$72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket (in dollars)?

128. State the distribution to use if you want to find the probability that the **mean** amount spent by five customers in one trip to the supermarket is less than \$60.

Chapter 13

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

129. What is the probability of both a drought **and** water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises:

	Apple	Pumpkin	Pecan
Female	40	10	30
Male	20	30	10

131. Suppose that one individual is randomly chosen. Find the probability that the person's favorite pie is apple **or** the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

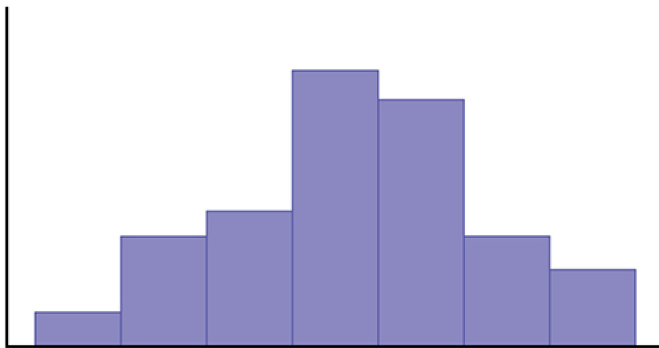
133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X . $X \sim$ _____

136. The following histogram is most likely to be a result of sampling from which distribution?



- a. Chi-Square
- b. Geometric
- c. Uniform
- d. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages (in years) as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint: Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	\bar{x}	s^2
Boys	55	82	29
Girls	60	86	46

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

- a. Maximize α and minimize β
- b. Minimize α and maximize β
- c. Maximize α and β
- d. Minimize α and β

Use the following information to answer the next three exercises: 120 people were surveyed as to their favorite beverage (non-alcoholic). The results are below.

Beverage/Age	0–9	10–19	20–29	30+	Totals
Milk	14	10	6	0	30
Soda	3	8	26	15	52

Beverage/Age	0–9	10–19	20–29	30+	Totals
Juice	7	12	12	7	38
Totals	24	330	44	22	120

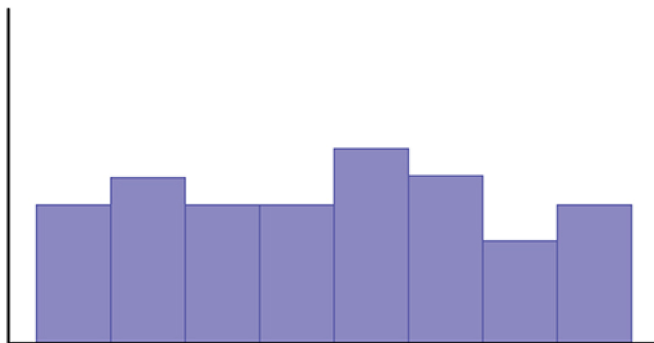
141. Are the events of milk and 30+:

- independent events? Justify your answer.
- mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are “Preferred Beverage” and “Age” independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?



- uniform
- exponential
- normal
- chi-square

Solutions

Chapter 3

1. c. parameter
2. a. population
3. b. statistic
4. d. sample
5. e. variable
6. quantitative continuous
7.
 - a. 2.27
 - b. 3.04
 - c. -1, 4, 4

8. Answers will vary.

Chapter 4

9. c. $(0.80)(0.30)$
10. b. No, and they are not mutually exclusive either.
11. a. all employed adult women
12. 0.5773
13. 0.0522
14. b. The middle fifty percent of the members lost from 2 to 8.5 lbs.
15. c. All of the data have the same value.

16. c. The lowest data value is the median.

17. 0.279

18. b. No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.

$X = 0, 1, 2, \dots, 25$

20. $B(25, 0.04)$

21. 0.0165

22. 1

23. c. quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

a. 24%

b. 27%

26. qualitative

27. 0.36

28. 0.7636

29.

a. No

b. No

30. $B(10, 0.76)$

31. 0.9330

32.

- a. X = the number of questions posted to the statistics listserv per day.
- b. $X = 0, 1, 2, \dots$
- c. $X \sim P(2)$
- d. 0

33. \$150

34. Matt

35.

- a. false
- b. true
- c. false
- d. false

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

- a. true
- b. true

- c. False – the median and the mean are the same for this symmetric distribution.
- d. true

42.

- a. 8
- b. 8
- c. $P(x < k) = 0.65 = (k - 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

- a. False – $\frac{3}{4}$ of the data are at most five.
- b. True – each quartile has 25% of the data.
- c. False – that is unknown.
- d. False – 50% of the data are four or less.

44. d. G and H are independent events.

45.

- a. False – J and K are independent so they are not mutually exclusive which would imply dependency (meaning $P(J \text{ AND } K)$ is not 0).
- b. False – see answer c.
- c. True – $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K) = P(J) + P(K) - P(J)P(K) = 0.3 + 0.6 - (0.3)(0.6) = 0.72$. Note the $P(J \text{ AND } K) = P(J)P(K)$ because J and K are independent.
- d. False – J and K are independent so $P(J) = P(J|K)$

46. a. $P(5)$

Chapter 7

47. a. $U(0, 4)$

48. b. 2 hour

49. a. $\frac{1}{4}$

50.

a. 0.7165

b. 4.16

c. 0

51. c. 5 years

52. c. exponential

53. 0.63

54. $B(14, 0.20)$

55. $B(14, 0.20)$

Chapter 8

56. c. the mean amount of weight lost by 15 people on the special weight loss diet.

57. 0.9951

58. 12.99

59. c. $\frac{1}{2}$

60. b. 0.60

61. c. $N(60, 5.477)$

62. 0.9990

63. a. eight days

64. c. 0.7500

65. a. 80%

66. b. 35%

67. b. no

68. b. quantitative continuous

69. c. 150

70. d. 0.06

71. c. 0.44

72. b. 0

Chapter 9

73. d. Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	0.3
2	0.2	0.2
4	0.4	0.4
5	0.1	0.1

76.

- a. 2.8
- b. 1.48
- c. 90%

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. d. $\frac{8}{70}$

80. c. $\frac{40}{70}$

81. a. $\frac{9}{19}$

82. b. false

83. b. false

84. b. false

85.

a. X = the number of pies Lee bakes every day.

b. $P(20)$

c. 0.1122

86. CI: (5.25, 8.48)

87.

a. uniform

b. exponential

c. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

- a. false
- b. false
- c. true
- d. false

91. $N(180, 16.43)$

92. a. The distribution for X is still uniform with the same mean and standard deviation as the distribution for X .

93. c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X .

94. $N\left(2, \frac{0.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. 0.5000

97. 7.6

98. 5

99. 0.9431

Chapter 11

100. 7.5

101. 0.0122

102. $N(7, 0.63)$

103. 0.9911

104. b. Exponential

105.

- a. true
- b. false
- c. false

106. Answers will vary.

107. Student's t with $df = 15$

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

- a. X = the number of patients with a shotgun wound the emergency room gets per 28 days
- b. $P(4)$
- c. 0.0183

112. greater than

113. No; $P(x = 8) = 0.0348$

114. You will lose \$5.

115. Becca

116. 14

117. Sample mean = 3.2

Sample standard deviation = 1.85

Median = 3

$Q_1 = 2$

$Q_3 = 5$

$IQR = 3$

118. d. $z = -1.19$

e. 0.1171

f. Do not reject the null hypothesis.

119. We conclude that the patient does have the HIV virus when, in fact, the patient does not.

120. c. $z = 2.21$; $p = 0.0136$

d. Reject the null hypothesis.

e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.

f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. c. dependent means

122. t_5

Chapter 12

123. (0.0424, 0.0770)

124. 2,401

125. Check student's solution.

126. 0.6321

127. \$360

128. $N\left(72, \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. 0.02

130. 0.40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. p -value = 0; Reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. $B(14, 0.60)$

136. d. Binomial

137. 0.3669

138. p -value = 0.0006; reject the null hypothesis; conclude that the averages are not equal

139. p -value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. Minimize α and β

141.

a. No

b. Yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. No; p -value = 0

144. a. uniform

References

Data from the *San Jose Mercury News*.

Baran, Daya. "20 Percent of Americans Have Never Used Email." Webguild.org, 2010. Available online at: <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email> (accessed October 17, 2013).

Data from *Parade Magazine*.

Practice Tests (1-4) and Final Exams

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.
 - a. population
 - b. sample
 - c. parameter
 - d. statistic
 - e. variable
 - f. data

2. What kind of data is "amount of money spent on produce per visit"?
 - a. qualitative
 - b. quantitative-continuous
 - c. quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is \$12.84. This is an example of a:
 - a. population
 - b. sample
 - c. parameter
 - d. statistic
 - e. variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?
 - a. cluster
 - b. stratified
 - c. simple random
 - d. systematic

5. “Number of visits per week” is what kind of data?

- a. qualitative
- b. quantitative-continuous
- c. quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question; that fifty percent is a:

- a. parameter
- b. statistic
- c. variable
- d. data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the U.S. concerning their plans for future education and employment, and found that 50 percent were planning to attend a 4-year college or university in the following year. This 50 percent is an example of a:

- a. parameter
- b. statistic
- c. variable
- d. data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25		
5–10	30		
> 10	empty		

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have ten or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

- a. cluster
- b. stratified
- c. simple random
- d. systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the NFL (National Football League) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, ages 3–18 months, from a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by ten percentage points over the previous year. Does this improvement prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:

14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups: the first is served a meal using normal-sized tableware, while the second is served the same meal, but using tableware that is 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

- a. population
- b. sample
- c. experimental units
- d. explanatory variable
- e. treatment
- f. response variable

23. A researcher analyzes the results of the SAT (Scholastic Aptitude Test) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of smoking.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were:

62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76, 78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range 0–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are:

1; 1; 1; 1

2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4; 4; 4; 4

5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80th percentile in math and the 76th percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82nd percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

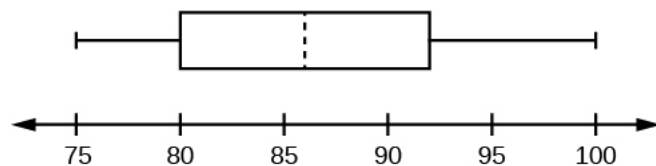
Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.



41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The value, in thousands of dollars, for houses on a block, are: 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.
48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?
- a. the mean
 - b. the media
 - c. the mode
50. In a right-skewed distribution, which is greater?
- a. the mean
 - b. the median
 - c. the mode
51. In a symmetrical distribution what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.
53. What number is two standard deviations above the mean of this data?
54. Express the number 13.7 in terms of the mean and standard deviation of this data.
55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85, and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a z-score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial, and replace it before the next trial. Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.
Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.
57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.

Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.
59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.
61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation:

$$P(A) = P(B) = P(\text{hitting the center on one shot}) = 0.70$$

$$P(B|A) = P(\text{hitting the center on a second shot, given that she hit it on the first}) = 0.85$$

63. Calculate the probability that she will hit the center of the target on two consecutive shots.
64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor roll	No honor roll	Total
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

65. Complete the table.
66. Find $P(\text{honor roll}|\text{study at least 15 hours per week})$.
67. What is the probability a student studies less than 15 hours per week?

68. Are the events “study at least 15 hours per week” and “makes the honor roll” independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team, some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

- a. population: all the shopping visits by all the store’s customers
- b. sample: the 1,000 visits drawn for the study
- c. parameter: the average expenditure on produce per visit by all the store’s customers
- d. statistic: the average expenditure on produce per visit by the sample of 1,000
- e. variable: the expenditure on produce for each visit
- f. data: the dollar amounts spent on produce; for instance, \$15.40, \$11.53, etc

2. c

3. d

1.2: Data, Sampling, and Variation in Data and Sampling

4. d

5. c

6. Answers will vary.

Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor might calculate the average exam score for her class: because the scores of all members of the class were used in the calculation, the average is a parameter.

7. b

8. a

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
5–10	30	0.30	0.55
> 10	45	0.45	1.00

10. 0.75

11. 0.55

12. Answers will vary.

Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then you would draw a proportionate random sample from within each class (for instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class).

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. a

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds, and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

- a. population: all college students
- b. sample: the 100 college students in the study
- c. experimental units: each individual college student who participated
- d. explanatory variable: the size of the tableware
- e. treatment: tableware that is 20 percent smaller than normal
- f. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either smoke or not smoke. Because smoking has many harmful effects, this would not be an ethical experiment. Instead, we study people who have chosen to smoke, and compare them to others who have chosen not to smoke, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of 0 to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8

Stem	Leaf
10	0 0

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5, central point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. $IQR = 92 - 80 = 12$

44. $\text{Range} = 100 - 75 = 25$

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or \$61,500

47. 49.25 or \$49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. c

50. a

51. They will all be fairly close to each other.

2.7: Measures of the Spread of the Data

52. Mean: 15

Standard deviation: 4.3

$$\mu = \frac{10+11+15+15+17+22}{6} = 15$$

$$s = \sqrt{\frac{\sum (x - \mu)^2}{n-1}} = \sqrt{\frac{94}{5}} = 4.3$$

53. $15 + (2)(4.3) = 23.6$

54. 13.7 is one standard deviation below the mean of this data, because $15 - 4.3 = 10.7$

$$55. z = \frac{95-85}{5} = 2.0$$

Susan's z-score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

$$56. P(B) = \frac{25}{90} = 0.28$$

57. Drawing a red marble is more likely.

$$P(R) = \frac{50}{80} = 0.62$$

$$P(Y) = \frac{15}{80} = 0.19$$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

$$60. P(A \text{ AND } B) = (0.3)(0.5) = 0.15$$

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	No honor roll	Total
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

66. $P(\text{honor roll} | \text{study at least 15 hours word per week}) = \frac{482}{1000} = 0.482$

67. $P(\text{studies less than 15 hours word per week}) = \frac{125+193}{1000} = 0.318$

68. Let $P(S)$ = study at least 15 hours per week

Let $P(H)$ = makes the honor roll

From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.

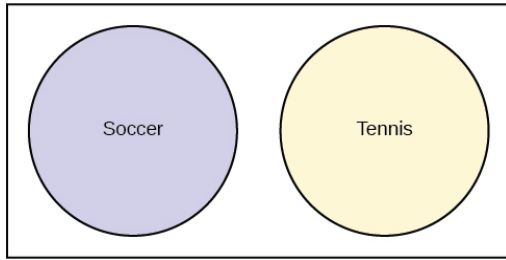
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.

However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.

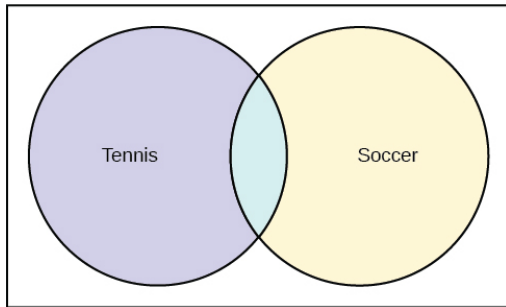
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.



70.



Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X ?
2. If Y = the number of classes taken in the previous semester, what is the domain of Y ?
3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z ?
4. Why are X , Y , and Z in the previous example random variables?
5. After collecting data, you find that for one case, $z = -7$. Is this a possible value for Z ?
6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	$P(x)$
-----	--------

x	$P(x)$
0	0.20
1	0.45
2	0.20
3	0.10
4	0.05

7. Define the random variable X for this example.
8. What is $P(x > 2)$?
9. What is the probability that a patron will check out at least one book?
10. What is the probability a patron will take out no more than three books?
11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?
12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in [\[link\]](#).

x	$P(x)$	$xP(x)$

14. What is the mean number of applicants?
15. What is the PDF for X ?
16. Add a fourth column to the table, for $(x - \mu)^2 P(x)$.
17. What is the standard deviation of X ?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particularly community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X .

22. What is the mean of X ?

23. What is the standard deviation of X ?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every ten times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X .

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X ?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X ?

33. What is the mean of X ?

34. What is the standard deviation of X ?

4.5: Hypergeometric Distribution

35. You draw a random sample of ten students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X ?

40. What are the mean and standard deviation of X ?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X ?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?
51. What is the probability that a patient waits less than ten minutes?

5.3: The Exponential Distribution

52. The distribution of the variable X , representing the average time to failure for an automobile battery, can be written as: $X \sim \text{Exp}(m)$. Describe this distribution in words.
53. If the value of m for an exponential distribution is ten, what are the mean and standard deviation for the distribution?
54. Write the probability density function for a variable distributed as: $X \sim \text{Exp}(0.2)$.

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable X into words: $X \sim (100, 15)$.
56. If the variable X has the standard normal distribution, express this symbolically.
- Use the following information for the next six exercises.* According to the World Health Organization, distribution of height in centimeters for girls aged five years and no months has the distribution: $X \sim N(109, 4.5)$.
57. What is the z-score for a height of 112 inches?
58. What is the z-score for a height of 100 centimeters?
59. Find the z-score for a height of 105 centimeters and explain what that means In the context of the population.
60. What height corresponds to a z-score of 1.5 in this population?
61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?
62. Using the empirical rule, about what percent of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of lotto tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.
64. What are the expected mean and standard deviation for your sample, assuming the distributor's claim is true?
65. What is the probability that your sample will have a mean greater than 100?
66. If the z-score for your sample result is -2.00 , explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. X is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to “the standard error of the mean.” What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the z -score for a sample mean of 76?

76. What is the z -score for a sample mean of 74.7?

77. What sample mean corresponds to a z -score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

- 84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?
- 85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?
- 86. What value is one standard deviation below the mean?
- 87. What value corresponds to a z-score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would expect to be closer to the population mean, a sample of size ten or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

- 90. If X = the diameter of one screw, what is the distribution of X ?
- 91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?
- 92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

- 1. The domain of $X = \{\text{English, Mathematics, ...}\}$, i.e., a list of all the majors offered at the university, plus “undeclared.”
- 2. The domain of $Y = \{0, 1, 2, \dots\}$, i.e., the integers from 0 to the upper limit of classes allowed by the university.
- 3. The domain of Z = any amount of money from 0 upwards.
- 4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.
- 5. No, because the domain of Z includes only positive numbers (you can’t spend a negative amount of money). Possibly the value -7 is a data entry error, or a special code to indicated that the student did not answer the question.
- 6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.
- 7. Let X = the number of books checked out by a patron.
- 8. $P(x > 2) = 0.10 + 0.05 = 0.15$
- 9. $P(x \geq 0) = 1 - 0.20 = 0.80$

10. $P(x \leq 3) = 1 - 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. $x = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35$

Mean or Expected Value and Standard Deviation

13.

x	$P(x)$	$xP(x)$
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

14. $x = 9.90 + 13.20 + 19.80 = 42.90$

15. $P(x = 30) = 0.33$

$P(x = 40) = 0.33$

$P(x = 60) = 0.33$

16.

x	$P(x)$	$xP(x)$	$(x - \mu)^2 P(x)$
30	0.33	9.90	$(30 - 42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40 - 42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60 - 42.90)^2(0.33) = 96.49$

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 - 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to 1.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. $X \sim B(100, 0.65)$

22. $\mu = np = 100(0.65) = 65$

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. X = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. $X \sim B(20, 0.4)$

26. $\mu = np = 20(0.4) = 8$

27. $\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. $T T T T H$

30. The domain of $X = \{1, 2, 3, 4, 5, \dots, n\}$. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \dots, 27\}$. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 27 draws.

32. $X \sim G(0.24)$

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of $X = \{0, 1, 2, 3, \dots\}$ i.e., any integer from 0 upwards.

40. $\mu = 4$
 $\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$, i.e., the probability that the value of x is between the values c and d .

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 - 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P(0 < x < 4) = (4 - 0) \left(\frac{1}{10}\right) = 0.4$

5.2: The Uniform Distribution

47. $P(2 < x < 5) = (5 - 2) \left(\frac{1}{10}\right) = 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b-a}$ for $(a \leq x \leq b)$ so $f(x) = \frac{1}{30}$ for $(0 \leq x \leq 30)$

50. $\mu = \frac{a+b}{2} = \frac{0+30}{2} = 15.0$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(30-0)^2}{12}} = 8.66$$

51. $P(x < 10) = (10) \left(\frac{1}{30}\right) = 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be a relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \geq 0$.

6.1: The Standard Normal Distribution

55. The random variable X has a normal distribution with a mean of 100 and a standard deviation of 15.

56. $X \sim N(0,1)$

57. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{112-109}{4.5} = 0.67$

58. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{100-109}{4.5} = -2.00$

59. $z = \frac{105-109}{4.5} = -0.89$

This girl is shorter than average for her age, by 0.89 standard deviations.

60. $109 + (1.5)(4.5) = 115.75$ cm

61. We expect about 68 percent of the heights of girls of age five years and zero months to be between 104.5 cm and 113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.

$np = (500)(0.20) = 100$ and $nq = 500(0.80) = 400$

64. $\mu = np = (500)(0.20) = 100$

$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the lotto tickets are winners, as claimed by the distributor, and that the true percent of winners is lower. Applying the Empirical Rule, If that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem states that for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $X \sim N(25, 0.2)$ because $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x , the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of n .

$$75. z = \frac{x - \mu_x}{\sigma_x} = \frac{76 - 75}{4.5} = 2.2$$

$$76. z = \frac{x - \mu_x}{\sigma_x} = \frac{74.7 - 75}{4.5} = -0.67$$

$$77. 75 + (1.5)(0.45) = 75.675$$

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size $n = 50$ is:

$$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:

$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable X , the random variable ΣX will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution, as the sample size increases, regardless of the distribution of population from which the samples are drawn.

$$83. \Sigma X \sim N(n\mu_x, (\sqrt{n})(\sigma_x)) \text{ so } \Sigma X \sim N(4000, 28.3)$$

84. The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum: $(\sqrt{n})(\sigma_x) = (\sqrt{40})(7) = 44.3$ so you would expect 95 percent of the values to be between $5,000 + (2)(44.3)$ and $5,000 - (2)(44.3)$, or between 4,911.4 and 5,088.6.

$$86. \mu - (\sqrt{n})(\sigma_x) = 5000 - (\sqrt{40})(7) = 4955.7$$

$$87. 5000 + (2.2) \left(\sqrt{40} \right) (7) = 5097.4$$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that as sample size increases, the sample mean tends to approach the population mean.

$$90. X \sim N(0.10, 0.20)$$

91. $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}} = \frac{0.10}{\sqrt{12}} = 0.03$
so $X \sim N(0.15, 0.003)$

$$92. \Sigma X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x)) \text{ so } \Sigma X \sim N(9.0, 0.23)$$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?
2. What is the distribution of the sample mean?
3. If you want to construct a two-sided 95% confidence interval, how much probability will be in each tail of the distribution?
4. What is the appropriate z-score and error bound or margin of error (*EBM*) for a 95% confidence interval for this data?
5. Rounding to two decimal places, what is the 95% confidence interval if the sample mean is 41?
6. What is the 90% confidence interval if the sample mean is 41? Round to two decimal places
7. Suppose the sample size in this study had been 50, rather than 30. What would the 95% confidence interval be if the sample mean is 41? Round your answer to two decimal places.
8. For any given data set and sampling situation, which would you expect to be wider: a 95% confidence interval or a 99% confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (z-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how do they differ?
10. Comparing graphs of the standard normal distribution (z-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the t -distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What is the degrees of freedom (df) for this study?
12. For a two-tailed 95% confidence interval, what is the appropriate t -value to use in the formula?
13. What is the 95% confidence interval?
14. What is the 99% confidence interval? Round to two decimal places.
15. Suppose your sample size had been 30 rather than 20. What would the 95% confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. 280 say they do own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.
17. What is the 95% two-sided confidence interval? Round to four decimal places.
18. Calculate the 90% confidence interval. Round to four decimal places.
19. Calculate the 99% confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members age 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95% confidence interval will be within four percentage points (plus or minus) the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?
21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?
22. Suppose you wanted a 95% confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as Republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.
24. You believe that at least 58 percent of registered voters in a community are registered as Republicans. State the null and alternative hypotheses to test this.
25. The mean household value in a city is \$268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.
26. State the appropriate alternative hypothesis to this null hypothesis: $H_0: \mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: $H_0: p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis, the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with np and nq greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college, and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against measles, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records, to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p -value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a p -value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29". You believe children with a particular disease are smaller than

average, so you draw a sample of 20 children with this disease and find a mean height of 27.5" and a sample standard deviation of 1.5".

42. What are the null and alternative hypotheses for this study?
43. What distribution will you use to test your hypothesis, and why?
44. What is the test statistic and the p -value?
45. Based on your sample results, what is your decision?
46. Suppose the mean for your sample was 25.0. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using $\alpha = 0.05$. What is the level of significance for this study?
48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
 $H_0: \mu = 35.5$
 $H_a: \mu \neq 35.5$
Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
 $H_0: \mu \geq 35.5$
 $H_a: \mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?
51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?
53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?
54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year's instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years

55. Calculate Cohen's d and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples t -test with sample size ten in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what p -values will cause you to reject the null hypothesis?

58. You conduct an independent samples t -test with sample size 15 in each group, with the following hypotheses:

$$H_0: \mu \geq 110$$

$$H_a: \mu < 110$$

If $\alpha = 0.05$, what t -values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of \$530 with a sample standard deviation of \$20 and the 50 humanities students spent an average of \$380 with a sample standard deviation of \$15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if proportion of homes with cable television service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study?

64. If 65 households in Community A have cable service, and 78 households in community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people lose weight. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You

compare the results using a matched-pairs t -test, in which the data is {weight at conclusion – weight at start}. You believe that, on average, the participants will have lost weight after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that $x_d = -5$, $s_d = 6$, and $n = 30$ (pairs).

69. What are the degrees of freedom for this statistic?

70. Using $\alpha = 0.05$, what is your decision regarding the effectiveness of this program in causing weight loss? What is the conclusion?

71. What would it mean if the t -statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school's most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (O)	Expected (E)	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
Enrolled	145				
Not enrolled	55				

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the p -value. At the 5% significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve _____.

78. For a chi-square distribution with five degrees of freedom, the curve is _____.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	Cell = No
Freshman	100	150
Senior	200	50

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where O = observed and E = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the $\alpha = 0.5$ significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five by two table. What is the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a hypothesis test at the 5% significance level.

Ethnic Group \ Response Type	Favor	Oppose	No Opinion	Row Total
White / Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

85. In a test of homogeneity, what must be true about the expected value of each cell?
86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?
87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypothesis to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5%; A 95% confidence interval contains 95% of the probability, and excludes five percent, and the five percent excluded is split evenly between the upper and lower tails of the distribution.

4. z-score = 1.96; $EBM = z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right) = (1.96)(0.73) = 1.4308$

5. $41 \pm 1.43 = (39.57, 42.43)$; Using the calculator function Zinterval, answer is (40.74, 41.26). Answers differ due to rounding.

6. The z-value for a 90% confidence interval is 1.645, so $EBM = 1.645(0.73) = 1.20085$.
The 90% confidence interval is $41 \pm 1.20 = (39.80, 42.20)$.
The calculator function Zinterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is: $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0.57$

$EBM = z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right) = (1.96)(0.57) = 1.12$

The 95% confidence interval is $41 \pm 1.12 = (39.88, 42.12)$.
The calculator function Zinterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99% confidence interval, because it includes all but one percent of the distribution. The 95% confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's t

9. The t -distribution will have more probability in its tails ("thicker tails") and less probability near the mean of the distribution ("shorter in the center").

10. Both distributions are symmetrical and centered at zero.

11. $df = n - 1 = 20 - 1 = 19$

12. You can get the t -value from a probability table or a calculator. In this case, for a t -distribution with 19 degrees of freedom, and a 95% two-sided confidence interval, the value is 2.093, i.e., $t_{\frac{\alpha}{2}} = 2.093$. The calculator function is invT(0.975, 19).

$$13. EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = (2.093) \left(\frac{0.3}{\sqrt{20}} \right) = 0.140$$

$$98.4 \pm 0.14 = (98.26, 98.54).$$

The calculator function Tinterval answer is (98.26, 98.54).

$$14. t_{\frac{\alpha}{2}} = 2.861. \text{ The calculator function is invT}(0.995, 19).$$

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = (2.861) \left(\frac{0.3}{\sqrt{20}} \right) = 0.192$$

$$98.4 \pm 0.19 = (98.21, 98.59). \text{ The calculator function Tinterval answer is (98.21, 98.59).}$$

$$15. df = n - 1 = 30 - 1 = 29. t_{\frac{\alpha}{2}} = 2.045$$

$$EBM = z_t \left(\frac{s}{\sqrt{n}} \right) = (2.045) \left(\frac{0.3}{\sqrt{30}} \right) = 0.112$$

$$98.4 \pm 0.11 = (98.29, 98.51). \text{ The calculator function Tinterval answer is (98.29, 98.51).}$$

8.3: Confidence Interval for a Population Proportion

$$16. p' = \frac{280}{500} = 0.56$$

$$q' = 1 - p' = 1 - 0.56 = 0.44$$

$$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$$

$$17. \text{ Because you are using the normal approximation to the binomial, } z_{\frac{\alpha}{2}} = 1.96.$$

Calculate the error bound for the population (EBP):

$$EBP = z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 1.96 (0.222) = 0.0435$$

Calculate the 95% confidence interval:

$$0.56 \pm 0.0435 = (0.5165, 0.6035).$$

The calculator function 1-PropZint answer is (0.5165, 0.6035).

$$18. z_{\frac{\alpha}{2}} = 1.64$$

$$EBP = z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 1.64 (0.0222) = 0.0364$$

$$0.56 \pm 0.03 = (0.5236, 0.5964). \text{ The calculator function 1-PropZint answer is (0.5235, 0.5965)}$$

$$19. z_{\frac{\alpha}{2}} = 2.58$$

$$EBP = z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 2.58 (0.0222) = 0.0573$$

$$0.56 \pm 0.05 = (0.5127, 0.6173).$$

The calculator function 1-PropZint answer is (0.5028, 0.6172).

$$20. EBP = 0.04 \text{ (because } 4\% = 0.04)$$

$$z_{\frac{\alpha}{2}} = 1.96 \text{ for a 95\% confidence interval}$$

$$n = \frac{z^2 pq}{EBP^2} = \frac{1.96^2 (0.5)(0.5)}{0.04^2} = \frac{0.9604}{0.0016} = 600.25$$

You need 601 subjects (rounding upward from 600.25).

$$21. n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2 (0.6)(0.4)}{0.04^2} = \frac{0.9220}{0.0016} = 576.24$$

You need 577 subjects (rounding upward from 576.24).

$$22. n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2 (0.5)(0.5)}{0.03^2} = \frac{0.9604}{0.0009} = 1067.11$$

You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. $H_0: p = 0.58$

$H_a: p \neq 0.58$

24. $H_0: p \geq 0.58$

$H_a: p < 0.58$

25. $H_0: \mu \geq \$268,000$

$H_a: \mu < \$268,000$

26. $H_a: \mu \neq 107$

27. $H_a: p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = $1 - \beta = 1 - P(\text{Type II error})$.

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a ten percent probability of a Type I error, meaning that ten percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's t -test.

36. The normal distribution or z -test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24} . You use the t -distribution because you don't know the population standard deviation, and the degrees of freedom are 24 because $df = n - 1$.

39. $X \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$

Because you know the population standard deviation, and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$

41. Reject the null hypothesis, because $\alpha \geq p$.

42. $H_0: \mu \geq 29.0$

$H_a: \mu < 29.0$

43. t_{19} . Because you do not know the population standard deviation, use the t -distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is -4.4721 and the p -value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the p -value is almost zero using the calculator function TTEST so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. $H_0: p = 0.8$

$H_a: p \neq 0.8$

51. You will use the normal test for a single population proportion because np and nq are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

$$55. d = \frac{x_1 - x_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$$

This is a small effect size, because 0.375 falls between Cohen's small (0.2) and medium (0.5) effect sizes.

$$56. d = \frac{x_1 - x_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$$

The effect size is 0.625 . By Cohen's standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. p -value < 0.01 .

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110 .

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $X_1 - X_2$, i.e., the mean difference in amount spent on textbooks for the two groups.

60. $H_0: X_1 - X_2 \leq 0$

$H_a: X_1 - X_2 > 0$

This could also be written as:

$H_0: X_1 \leq X_2$

$H_a: X_1 > X_2$

61. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 1% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. $H_0: p_A = p_B$

$H_a: p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65 + 78}{100 + 100} = 0.715$

65. Using the calculator function 2-PropZTest, the p -value = 0.0417. Reject the null hypothesis. At the 3% significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the p -value = 0.0417. Do not reject the null hypothesis. At the 1% significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. $H_0: x_d \geq 0$

$H_a: x_d < 0$

68. $t = -4.5644$

69. $df = 30 - 1 = 29$.

70. Using the calculator function TTEST, the p -value = 0.00004 so reject the null hypothesis. At the 5% level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive t -statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$

$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = $200(0.66) = 132$. Not enrolled = $200(0.34) = 68$

74.

	Observed (O)	Expected (E)	O – E	(O – E) ²	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

75. $df = n - 1 = 2 - 1 = 1$.

76. Using the calculator function Chi-square GOF – Test (in STAT TESTS), the test statistic is 3.7656 and the p-value is 0.0523. Do not reject the null hypothesis. At the 5% significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	Cell = No	Total
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

80. $\frac{(100-150)^2}{150} = 16.67$

$\frac{(150-100)^2}{100} = 25$

$\frac{(200-100)^2}{150} = 16.67$

$\frac{(50-100)^2}{100} = 25$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.

$df = (r - 1)(c - 1) = 1$

82. $p\text{-value} = P(\text{Chi-square}, 83.34) = 0$

Reject the null hypothesis.

You could also use the calculator function STAT TESTS Chi-Square – Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. $df = (r - 1)(c - 1) = (4)(1) = 4$.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-square Test, the $p\text{-value} = 0$. Reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the poll responses independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0 : The variables are independent.

H_a : The variables are not independent.

87. H_0 : The populations have the same distribution.

H_a : The populations do not have the same distribution.

11.6: Test of a Single Variance

88. $H_0: \sigma^2 \leq 5$

$H_a: \sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

a. $y = -3x$

b. $y = 0.2 + 0.74x$

c. $y = -9.4 - 2x$

d. A and B

e. A, B, and C

2. To complete a painting job requires four hours setup time plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of \$2,000 plus \$100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of \$500 plus tuition of \$3,000 per year. Express this information in an equation.

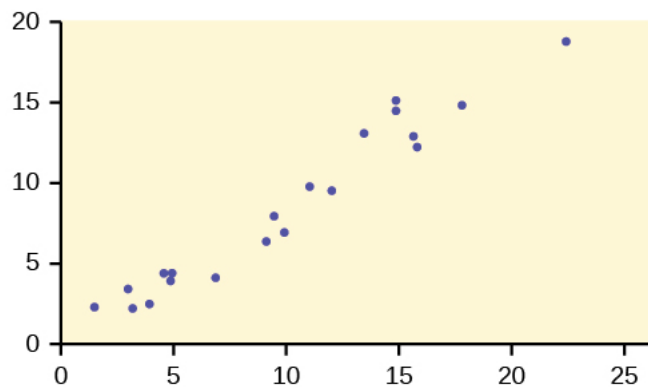
12.2: Slope and Y-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of \$75 per car, plus an hourly rate of \$55.

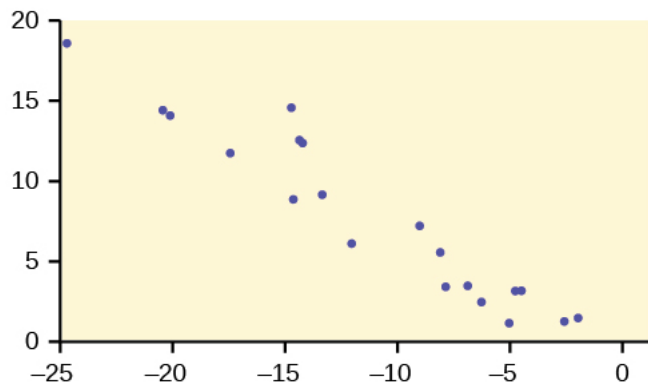
5. What are the independent and dependent variables for this situation?
6. Write the equation and identify the slope and intercept.
7. What is the labor charge for a job that takes 3.5 hours to complete?
8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

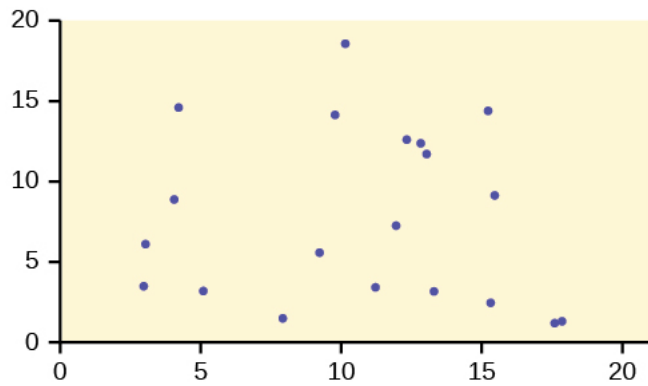
9. Describe the pattern in this scatter plot, and decide whether the X and Y variables would be good candidates for linear regression.



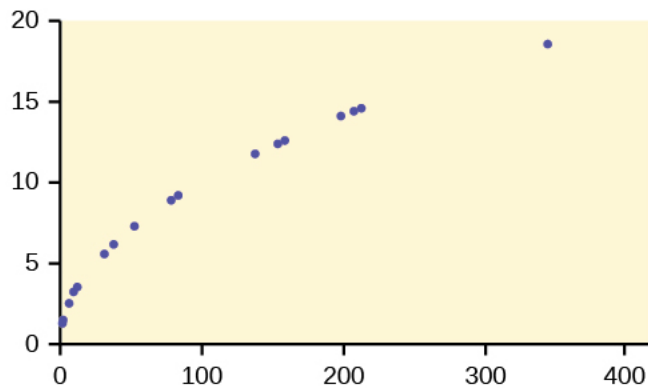
10. Describe the pattern in this scatter plot, and decide whether the X and Y variables would be good candidates for linear regression.



11. Describe the pattern in this scatter plot, and decide whether the X and Y variables would be good candidates for linear regression.



12. Describe the pattern in this scatter plot, and decide whether the X and Y variables would be good candidates for linear regression.



12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (In pounds) in a sample of college freshman men have a linear relationship with the following summary statistics:

$$\bar{x} = 68.4$$

$$\bar{y} = 141.6$$

$$s_x = 4.0$$

$$s_y = 9.6$$

$$r = 0.73$$

Let Y = weight and X = height, and write the regression equation in the form:

$$\hat{y} = a + bx$$

13. What is the value of the slope?

14. What is the value of the y intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is -0.56 . Calculate the coefficient of determination for this data and explain what it means.
17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32 . How much variation in freshman college GPA is not explained by high school GPA?
18. Rounded to two decimal places what correlation between two variables is necessary to have a coefficient of determination of at least 0.50 ?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.
20. In a sample of 30 cases, two variables have a correlation of 0.33 . Do a t -test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45 . Do a t -test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (Y) to the grams of fiber (X) per serving in enriched flour products (bread, rolls, etc.) produced the equation:

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?
23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?
25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?
26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4 . Does a case with the values $x = 2$, $y = 6.2$ qualify as an outlier?
27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13 . Does a case with the values $x = 4.1$, $y = 2.34$ qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples t -test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3 , for the samples taught by the different methods have the following distributions:

$$X_1 \sim N(85, 3.6)$$

$$X_2 \sim N(82, 4.8)$$

$$X_3 \sim N(79, 2.9)$$

Each sample includes 100 students, and the final exam scores have a range of 0–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The F Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are difference in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are k and i for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within} , for this experiment?

38. What is the F Statistic for this data?

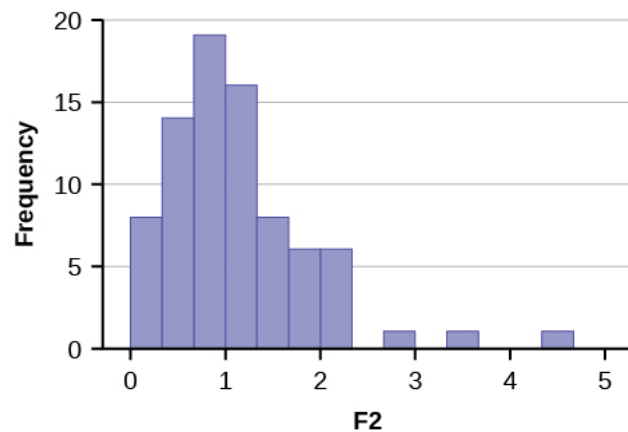
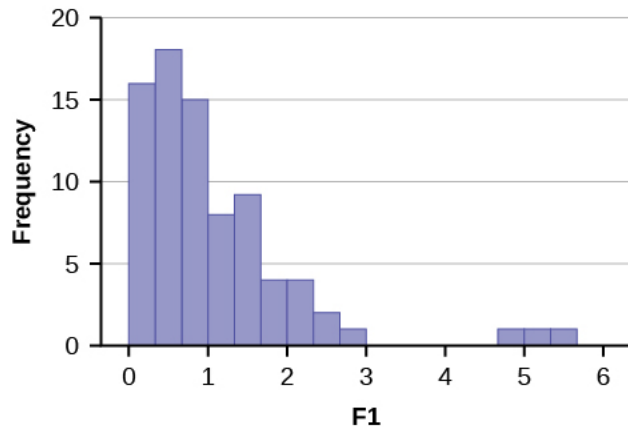
39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the F Statistic be larger or smaller?

13.3: Facts About the F Distribution

40. Which of the following numbers are possible F Statistics?

- a. 2.47
- b. 5.95
- c. -3.61
- d. 7.28
- e. 0.97

41. Histograms $F1$ and $F2$ below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?



42. The F Statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The F Statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the F test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, and compute the variance of each, and conduct an F test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.

All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is:
 $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is: $y = 100(x) + 2,000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is: $y = 3,000(x) + 500$

12.2: Slope and Y-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is: $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is: $6.3 - 2.4 = 3.9$. Multiply this difference by the cost per hour: $55(3.9) = 214.5$.
The difference in cost between the two jobs is \$214.50.

12.3: Scatter Plots

9. The X and Y variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The X and Y variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the X and Y variables, so they are not good candidates for linear regression.

12. The X and Y variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r \left(\frac{s_y}{s_x} \right) = 0.73 \left(\frac{9.6}{4.0} \right) = 1.752 \approx 1.75$

14. $a = y - bx = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2 .
For this data, $r^2 = (-0.56)^2 = 0.3136 \approx 0.31$ or 31%. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination $= 0.32^2 = 0.1024$. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is $1 - 0.1024 = 0.8976 \approx 0.90$. So about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$

$\sqrt{0.5} = 0.707106781 \approx 0.71$

You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. $H_0: \rho = 0$

$H_a: \rho \neq 0$

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

Using the calculator function tcdf, the p -value is $2\text{tcdf}(1.85, 10^{99}, 29) = 0.0373$. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

Using the calculator function tcdf, the p -value is $2\text{tcdf}(2.417, 10^{99}, 24) = 0.0118$. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is $6 - 3 = 3$ and the predicted difference in grams of potassium per serving is $(16)(3) = 48$.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is: $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.

Residual for (2, 6.2): $6.2 - 5.6 = 0.6$ ($0.6 < 2(0.4)$)

27. The predicted value for y is: $\hat{y} = 2.3 - 0.1(4.1) = 1.89$. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.
Residual for (4.1, 2.34): $2.32 - 1.89 = 0.43$ ($0.43 > 2(0.13)$)

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population
2. All samples are independent and randomly selected.
3. The populations from which the samples are drawn have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a : At least two of the group means $\mu_1, \mu_2, \mu_3, \mu_4$ are not equal.

30. The independent samples t -test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from a normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The F Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements, but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4

n_1 = the number of cases in group 1 = 30

n = the total number of cases = $4(30) = 120$

36. $SS_{total} = SS_{within} + SS_{between}$ so $SS_{between} = SS_{total} - SS_{within}$
 $621.4 - 374.5 = 246.9$

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (df).

For SS_{total} , $df = n - 1 = 120 - 1 = 119$.

For $SS_{between}$, $df = k - 1 = 4 - 1 = 3$.

For SS_{within} , $df = 120 - 4 = 116$.

$MS_{between} = \frac{246.9}{3} = 82.3$

$MS_{within} = \frac{374.5}{116} = 3.23$

$$38. F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{82.3}{3.23} = 25.48$$

39. It would be larger, because you would be dividing by a smaller number. The value of MS_{between} would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the F Distribution

40. All but choice c, -3.61 . F Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an F distribution, the distribution becomes more nearly normal. Histogram $F2$ is closer to a normal distribution than histogram $F1$, so the sample displayed in histogram $F1$ was drawn from the $F_{3,15}$ population, and the sample displayed in histogram $F2$ was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, $p\text{-value} = \text{Fcdf}(3.67, 1\text{E}, 3, 50) = 0.0182$. Reject the null hypothesis.

43. Using the calculator function Fcdf, $p\text{-value} = \text{Fcdf}(4.72, 1\text{E}, 4, 100) = 0.0016$. Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.

$$H_0: \sigma_M^2 \leq \sigma_E^2$$

$$H_a: \sigma_M^2 > \sigma_E^2$$

Practice Final Exam 1

Use the following information to answer the next two exercises: An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event A = both dice show an even number.
- Let Event B = both dice show a number more than eight

1. Events A and B are:

- a. mutually exclusive.
- b. independent.
- c. mutually exclusive and independent.
- d. neither mutually exclusive nor independent.

2. Find $P(A|B)$.

- a. $\frac{2}{4}$
- b. $\frac{16}{144}$
- c. $\frac{4}{16}$
- d. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

- a. Sample sizes are almost never small.
- b. Two measurements are drawn from the same pair of individuals or objects.

- c. Two sample means are compared to each other.
- d. Answer choices b and c are both true.

Use the following information to answer the next two exercises: One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.
- a. $\frac{10}{118}$
 - b. $\frac{68}{118}$
 - c. $\frac{48}{118}$
 - d. $\frac{10}{48}$
5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.
- a. $\frac{30}{118}$
 - b. $\frac{30}{48}$
 - c. $\frac{22}{118}$
 - d. $\frac{30}{52}$

Use the following information to answer the next two exercises: We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. [\[link\]](#) shows the results of the survey.

x	P (x)
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$

x	$P(x)$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

6. Find the probability that a teenager is reminded two times.

- a. 8
- b. $\frac{8}{40}$
- c. $\frac{6}{40}$
- d. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

- a. 15
- b. 2.78
- c. 1.0
- d. 3.13

Use the following information to answer the next two exercises: On any given day, approximately 37.5% of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

- a. 8.25
- b. 11
- c. 18
- d. 7.5

9. What is the probability that at least ten of the 22 cars are parked crookedly.

- a. 0.1263
- b. 0.1607
- c. 0.2870
- d. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. The correct distribution to use for the hypothesis test is:

- a. Binomial
- b. Student's t
- c. Normal
- d. Uniform

Use the following information to answer the next three exercises: De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be:

- a. $H_a: p_A = p_B$
- b. $H_a: p_A > p_B$
- c. $H_o: p_A = p_B$
- d. $H_a: p_A \neq p_B$

12. The Type I error is to:

- a. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.
- b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.
- c. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.
- d. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to:

- a. reject H_0
- b. not reject H_0
- c. There is not enough information given to conduct the hypothesis test

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

14. Which student is the BEST when compared to the other runners at her school?

- a. Kia
- b. Alejandra
- c. Iris
- d. Impossible to determine

Use the following information to answer the next two exercises: The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: \$212, \$292, \$278, \$199, \$280, \$236

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least \$275.

15. The correct distribution to use for the hypothesis test is:

- a. Normal
- b. Binomial

- c. Student's t
- d. Exponential

16. The hypothesis test:

- a. is two-tailed.
- b. is left-tailed.
- c. is right-tailed.
- d. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

- a. simple random sampling.
- b. systematic sampling.
- c. cluster sampling.
- d. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical "middle" customer?

- a. mean
- b. median
- c. mode
- d. IQR

19. In a hypothesis test, the p -value is

- a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
- b. called the preconceived alpha.
- c. compared to beta to decide whether to reject or not reject the null hypothesis.
- d. Answer choices A and B are both true.

Use the following information to answer the next three exercises: A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in [\[link\]](#).

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
6	1	.02	1.00

20. Combined with convenience sampling, what other sampling technique did Maria use?

- a. simple random
- b. systematic
- c. cluster
- d. stratified

21. How many students come to campus for classes four days a week?

- a. 49
- b. 25
- c. 30
- d. 13

22. What is the 60th percentile for the this data?

- a. 2
- b. 3
- c. 4
- d. 5

Use the following information to answer the next two exercises: The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

23. Construct a 95% confidence interval for the true population mean number of dependents of Reservists called to active duty to increase security at California airports.

- a. (1.85, 2.32)
- b. (1.80, 2.36)
- c. (1.97, 2.46)
- d. (1.92, 2.50)

24. The 95% confidence interval above means:

- a. Five percent of confidence intervals constructed this way will not contain the true population average number of dependents.
- b. We are 95% confident the true population mean number of dependents falls in the interval.
- c. Both of the above answer choices are correct.
- d. None of the above.

25. $X \sim U(4, 10)$. Find the 30th percentile.

- a. 0.3000
- b. 3
- c. 5.8
- d. 6.1

26. If $X \sim \text{Exp}(0.8)$, then $P(x < \mu) = \underline{\hspace{2cm}}$

- a. 0.3679
- b. 0.4727
- c. 0.6321
- d. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

- a. 0.8413
- b. 0.1587
- c. 0.3461
- d. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98% confidence interval for the true population proportion of reservists called to active duty that are married.

- a. (0.6030, 0.7954)
- b. (0.6181, 0.7802)
- c. (0.5927, 0.8057)
- d. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last ten marathon winning times is collected. Let x = mean winning times for ten marathons. The distribution for x is:

- a. $N\left(145, \frac{14}{\sqrt{10}}\right)$
- b. $N(145, 14)$
- c. t_9
- d. t_{10}

30. Suppose that Phi Beta Kappa honors the top one percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

- a. 3.99
- b. 1.34
- c. 3.00
- d. 3.66

The number of people living on American farms has declined steadily during the 20th century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

31. The linear regression equation is $\hat{y} = 1166.93 - 0.5868x$. What was the expected farm population (in millions of persons) for 1980?

- a. 7.2
- b. 5.1
- c. 6.0
- d. 8.0

32. In linear regression, which is the best possible SSE?

- a. 13.46
- b. 18.22
- c. 24.05
- d. 16.33

33. In regression analysis, if the correlation coefficient is close to one what can be said about the best fit line?

- a. It is a horizontal line. Therefore, we can not use it.
- b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.
- c. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
- d. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises: A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is:

- a. Chi^2_8 .
- b. Chi^2_3 .
- c. t_{721} .
- d. $N(0, 1)$.

35. The expected number of female who choose finance is:

- a. 37.
- b. 61.
- c. 60.
- d. 70.

36. The p -value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

- a. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
- b. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
- c. there is sufficient evidence to conclude that students find economics very hard.
- d. there is sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10% professional, 10% clerical, 30% skilled, 15% service, and 35% semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$ does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

- a. χ^2 goodness of fit
- b. χ^2 test of independence
- c. Independent groups proportions
- d. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. b. independent

2. c. $\frac{4}{16}$

3. b. Two measurements are drawn from the same pair of individuals or objects.

4. b. $\frac{68}{118}$

5. d. $\frac{30}{52}$

6. b. $\frac{8}{40}$

7. b. 2.78

8. a. 8.25

9. c. 0.2870

10. c. Normal

11. d. $H_a: p_A \neq p_B$

12. b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. b. not reject H_0

- 14. c. Iris
- 15. c. Student's t
- 16. b. is left-tailed.
- 17. c. cluster sampling
- 18. b. median
- 19. a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
- 20. d. stratified
- 21. b. 25
- 22. c. 4
- 23. a. (1.85, 2.32)
- 24. c. Both above are correct.
- 25. c. 5.8
- 26. c. 0.6321
- 27. a. 0.8413
- 28. a. (0.6030, 0.7954)
- 29. a. $N\left(145, \frac{14}{\sqrt{10}}\right)$
- 30. d. 3.66
- 31. b. 5.1
- 32. a. 13.46
- 33. b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.
- 34. b. Chi^2_3 .
- 35. d. 70
- 36. b. There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
- 37. a. Chi^2 goodness-of-fit

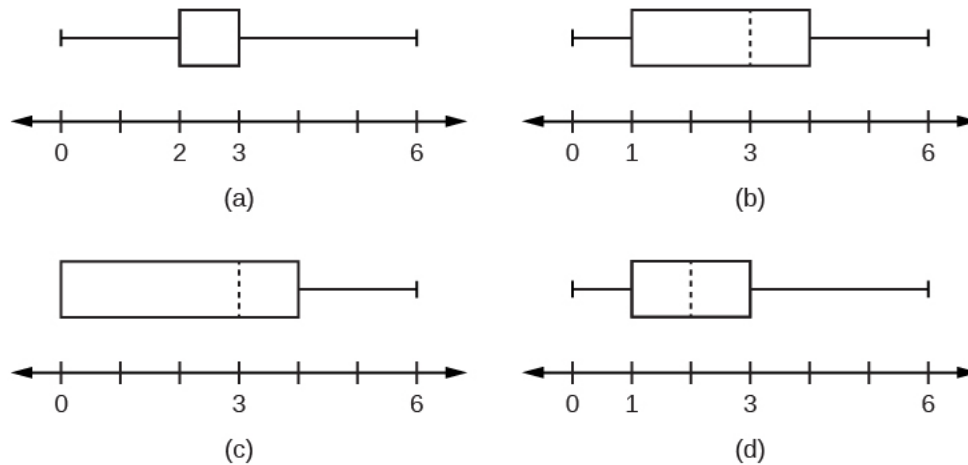
Practice Final Exam 2

1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the:
- a. statistic.
 - b. parameter.
 - c. population.
 - d. variable.

Use the following information to answer the next two exercises:

value	frequency
0	1
1	4
2	7
3	9
6	4

2. The box plot for the data is:



3. If six were added to each value of the data in the table, the 15th percentile of the new list of values is:

- a. six
- b. one
- c. seven
- d. eight

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is ten percent. However, in any year, the probability of water rationing is five percent.

4. What is the probability of both a drought and water rationing occurring?

- a. 0.05
- b. 0.01

- c. 0.02
- d. 0.30

5. Which of the following is true?

- a. Drought and water rationing are independent events.
- b. Drought and water rationing are mutually exclusive events.
- c. None of the above

Use the following information to answer the next two exercises: Suppose that a survey yielded the following data:

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is _____.

- a. $\frac{40}{60}$
- b. $\frac{60}{140}$
- c. $\frac{120}{140}$
- d. $\frac{100}{140}$

7. Suppose H_0 is: Favorite pie and gender are independent. The p -value is _____.

- a. ≈ 0
- b. 1
- c. 0.05
- d. cannot be determined

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

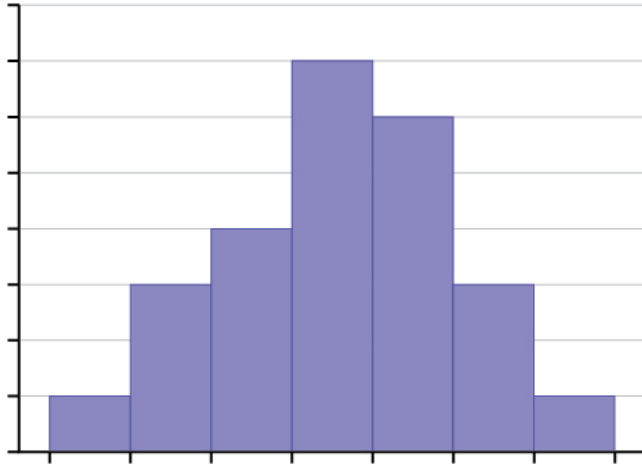
- a. $X \sim B(14, 0.60)$
- b. The values for x are: $\{1, 2, 3, \dots, 14\}$.
- c. $\mu = 8.4$
- d. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

- a. $\frac{6}{14}$
- b. 0.8499
- c. 0.9417

d. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?



- a. chi-square with $df = 6$
- b. exponential
- c. uniform
- d. binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}. What is the error bound for the 90% confidence interval of the true average age?

- a. 11.2
- b. 22.3
- c. 17.5
- d. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5% of the population values lie above:

- a. -1.96.
- b. 1.96.
- c. 1.
- d. -1.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is \$72.

13. What is the probability that one customer spends less than \$72 in one trip to the supermarket?

- a. 0.6321
- b. 0.5000
- c. 0.3714
- d. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

- a. 72
- b. $\frac{72^2}{5}$
- c. 5184
- d. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than \$60, the distribution to use is:

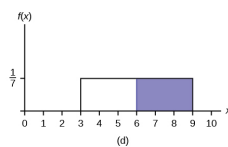
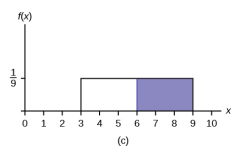
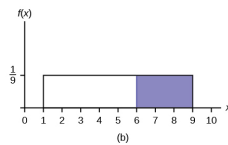
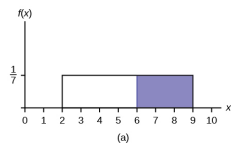
- a. $N(72, 72)$
- b. $N\left(72, \frac{72}{\sqrt{50}}\right)$
- c. $Exp(72)$
- d. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises: The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to ten minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

- a. $\frac{3}{9}$
- b. $\frac{7}{9}$
- c. $\frac{3}{10}$
- d. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash given that he or she has already taken more than three minutes?



18. We should expect a fourth grader to take how many minutes to take out the trash?

- a. 4.5
- b. 5.5
- c. 5
- d. 10

Use the following information to answer the next three exercises: At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90th percentile of waiting times (in minutes)?

- a. 1.28

- b. 90
- c. 7.47
- d. 6.92

20. The median waiting time (in minutes) for one student is:

- a. 5.
- b. 50.
- c. 2.5.
- d. 1.5.

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

- a. 0.6301
- b. 0.8541
- c. 0.3694
- d. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken and it is found that 20% of them earn approximately \$50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn \$50,000 per year is:

- a. 16.
- b. 0.2.
- c. 1.
- d. 0.95.

23. If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to:

- a. -1.
- b. 0.1587.
- c. 0.8413.
- d. 1.

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of:

- a. independent groups.
- b. two proportions.
- c. matched pairs, dependent groups.
- d. exclusive groups.

A math exam was given to all the third grade children attending ABC School. Two random samples of scores were taken.

	<i>n</i>	<i>x</i>	<i>s</i>
Boys	55	82	5
Girls	60	86	7

25. Which of the following correctly describes the results of a hypothesis test of the claim, “There is a difference between the mean scores obtained by third grade girls and boys at the 5% level of significance”?

- a. Do not reject H_0 . There is insufficient evidence to conclude that there is a difference in the mean scores.
- b. Do not reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.
- c. Reject H_0 . There is insufficient evidence to conclude that there is no difference in the mean scores.
- d. Reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that:

- a. there is insufficient information to conclude that the proportion for males is the same as the proportion for females.
- b. there is insufficient information to conclude that the proportion for males is not the same as the proportion for females.
- c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
- d. not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81 with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher’s claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

- a. Yes
- b. No
- c. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages with a standard deviation of one minute. Find the p -value when testing that the average loading times are the same.

- a. p -value is close to zero
- b. p -value is close to one
- c. not enough information given to solve the problem

Use the following information to answer the next three exercises: A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees x	650	730	810	900	102	107	1150
Number of bathrooms y	40	50	54	61	82	110	121

29. Is the correlation between the number of employees and the number of bathrooms significant?

- a. Yes
- b. No
- c. Not enough information to answer question

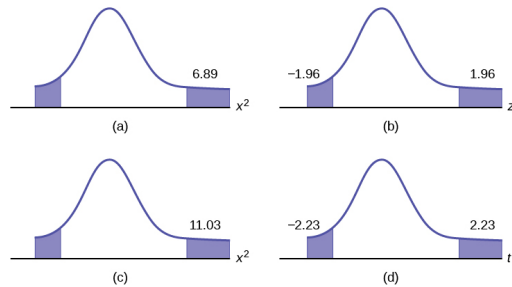
30. The linear regression equation is:

- a. $\hat{y} = 0.0094 - 79.96x$
- b. $\hat{y} = 79.96 + 0.0094x$
- c. $\hat{y} = 79.96 - 0.0094x$
- d. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

- a. 69
- b. 91
- c. 91,954
- d. We should not be estimating here.

32. Suppose that a sample of size ten was collected, with $x = 4.4$ and $s = 1.4$. $H_0: \sigma^2 = 1.6$ vs. $H_a: \sigma^2 \neq 1.6$. Which graph best describes the results of the test?



Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in [\[link\]](#):

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

33. Conduct an appropriate test to determine if the distribution is uniform.

- a. The p -value is > 0.10 . There is insufficient information to conclude that the distribution is not uniform.
- b. The p -value is < 0.01 . There is sufficient information to conclude the distribution is not uniform.
- c. The p -value is between 0.01 and 0.10, but without alpha (α) there is not enough information
- d. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

- a. The populations from which the samples are selected have different distributions.
- b. The sample sizes are large.
- c. The test is to determine if the different groups have the same means.
- d. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. b. parameter.
2. a.
3. c. seven
4. c. 0.02
5. c. none of the above
6. d. $\frac{100}{140}$
7. a. ≈ 0
8. b. The values for x are: $\{1, 2, 3, \dots, 14\}$
9. c. 0.9417.
10. d. binomial
11. d. 8.7
12. a. -1.96
13. a. 0.6321
14. d. 360
15. b. $N\left(72, \frac{72}{\sqrt{50}}\right)$
16. a. $\frac{3}{9}$
17. d.
18. b. 5.5
19. d. 6.92
20. a. 5
21. b. 0.8541
22. b. 0.2
23. a. -1 .
24. c. matched pairs, dependent groups.
25. d. Reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.
26. c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
27. b. no
28. b. p -value is close to 1.

29. b. No

30. c. $\hat{y} = 79.96x - 0.0094$

31. d. We should not be estimating here.

32. a.

33. a. The p -value is > 0.10 . There is insufficient information to conclude that the distribution is not uniform.

34. c. The test is to determine if the different groups have the same means.

Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Race Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

\$17.00	\$23.00	\$14.00	\$16.00	\$12.00	\$26.00
\$20.00	\$22.00	\$14.00	\$15.00	\$22.00	\$18.00
\$18.00	\$21.00	\$21.00	\$19.00	\$15.00	\$21.00
\$18.00	\$17.00	\$15.00	\$25.00	\$14.00	\$30.00
\$16.00	\$10.00	\$20.00	\$12.00	\$16.00	\$17.44

\$16.00	\$14.00	\$15.00	\$20.00	\$20.00	\$16.00
\$17.00	\$16.00	\$15.00	\$15.00	\$19.00	\$48.00
\$16.00	\$18.00	\$9.00	\$18.00	\$18.00	\$20.00
\$8.00	\$20.00	\$17.00	\$14.00	\$11.00	\$16.00
\$19.00	\$15.00	\$21.00	\$12.00	\$8.00	\$16.00
\$13.00	\$14.00	\$15.00	\$14.00	\$13.41	\$28.00
\$21.00	\$17.00	\$28.00	\$17.00	\$19.00	\$16.00
\$17.00	\$19.00	\$18.00	\$17.00	\$15.00	
\$14.00	\$21.00	\$12.00	\$18.00	\$24.00	
\$15.00	\$23.00	\$14.00	\$16.00	\$12.00	
\$24.00	\$20.00	\$14.00	\$14.00	\$15.00	
\$14.00	\$19.00	\$16.00	\$38.00	\$20.00	
\$24.00	\$16.00	\$8.00	\$18.00	\$17.00	
\$16.00	\$15.00	\$7.00	\$19.00	\$12.00	
\$8.00	\$23.00	\$12.00	\$18.00	\$20.00	
\$21.00	\$34.00	\$16.00	\$26.00	\$14.00	

IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

Group and Partner Projects

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

_____ Decide what data you are going to study.

Note:

Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

_____ Are your data discrete or continuous? How do you know?

_____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).

_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?

_____ Conduct your survey. **Your data size must be at least 30.**

_____ Summarize your data in a chart with columns showing **data value**,

frequency, relative frequency and cumulative relative frequency.

Answer the following (rounded to two decimal places):

- a. \bar{x} = _____
- b. s = _____
- c. First quartile = _____
- d. Median = _____
- e. 70th percentile = _____

_____ What value is two standard deviations above the mean?

_____ What value is 1.5 standard deviations below the mean?

_____ Construct a histogram displaying your data.

_____ In complete sentences, describe the shape of your graph.

_____ Do you notice any potential outliers? If so, what values are they?

Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.

_____ Construct a box plot displaying your data.

_____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.

_____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- _____ **Cover sheet:** name, class time, and name of your study
- _____ **Summary page:** This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- _____ **URL** for data, if your data are from the World Wide Web

- ____ Chart of data, frequency, relative frequency, and cumulative relative frequency
- ____ Page(s) of graphs: histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)

____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?

____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data.**

____ Define (in words) the random variable for your data. $X =$ _____

____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.

_____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. \bar{x} = _____

b. s = _____

_____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

_____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.

_____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b .

_____ Normal: $X \sim N$ _____ Use \bar{x} to estimate for μ and s to estimate for σ .

_____ **Must** your data fit one of the above distributions? Explain why or why not.

_____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.

_____ Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to two decimal places)

Note: $k = \bar{x} + (1.75)s$

_____ Determine the relative frequencies (RF) rounded to four decimal places.

Note:

Note

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

- a. $RF(X < k) =$ _____
- b. $RF(X > k) =$ _____
- c. $RF(X = k) =$ _____

Note:

Note

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

_____ State the distribution: $X \sim$ _____

_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.

_____ Find the following theoretical probabilities (rounded to four decimal places).

- a. $P(X < k) =$ _____
- b. $P(X > k) =$ _____
- c. $P(X = k) =$ _____

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?

_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.

_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.

_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, X , in words. $X =$ _____
_____ State the approximate theoretical distribution of X . $X \sim$ _____

_____ Base this on the mean and standard deviation from your original data.

_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.

Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. $k =$ _____ (rounded to two decimal places)

Determine the relative frequencies (RF) rounded to four decimal places.

a. $RF(X < k) =$ _____

b. $RF(X > k) =$ _____

c. $RF(X = k) =$ _____

Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k) =$ _____

b. $P(X > k) =$ _____

c. $P(X = k) =$ _____

_____ Draw the graph of the theoretical distribution of X .

_____ Compare the relative frequencies to the probabilities. Are the values close?

_____ Does it appear that the data of averages fit the distribution of X well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.

In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.

_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.

_____ What happened to the shape and distribution when you averaged your data? **In theory**, what should have happened? In theory, would “it”

always happen? Why or why not?

_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or X distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

_____ **Cover sheet:** name, class time, and name of your study

_____ **Summary pages:** These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions

_____ **URL** for data, if your data are from the World Wide Web

_____ **Pages, one for each theoretical distribution,** with the distribution stated, the graph, and the probability questions answered

_____ **Pages of the data requested**

_____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.

____ **Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.

____ **Copy or print out the article** and include a copy in your project, along with the source.

____ **State how you will collect your data.** (Convenience sampling is not acceptable.)

____ **Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.

____ **State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

____ **Make two copies of the appropriate solution sheet.**

____ **Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

____ **Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.

____ **Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

- a. Brief discussion of the article, including the source
- b. Statement of the claim made in the article (one of the hypotheses).
- c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified,

- systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
- d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
 - e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

____ **Cover sheet** containing your name(s), class time, and the name of your study

____ **Summary**, which includes all items listed on summary checklist

____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.

____ **Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

____ **Raw data collected AND a table summarizing the sample data** (n , x and s ; or x , n , and p' , as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.

- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

_____ State the bivariate data your group is going to study.

Note:

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.

_____ Conduct your survey. Your number of pairs must be at least 30.

_____ Print out a copy of your data.

Analysis

_____ On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.

_____ State the least squares line and the correlation coefficient.

_____ On your scatter plot, in a different color, construct the least squares line.

_____ Is the correlation coefficient significant? Explain and show how you determined this.

_____ Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.

_____ Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.

_____ Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____ Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.

_____ Answer the following question, rounded to two decimal places:

- a. Sample mean = _____
- b. Sample standard deviation = _____
- c. First quartile = _____
- d. Third quartile = _____
- e. Median = _____

- f. 70th percentile = _____
- g. Value that is 2 standard deviations above the mean = _____
- h. Value that is 1.5 standard deviations below the mean = _____

_____ Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26, 27-33, 34-40, 41-47, 48-54, 55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .

_____ In complete sentences, describe the shape of your histogram.

_____ Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in [Descriptive Statistics](#) (since you are now using univariate data) to determine which values might be outliers.

_____ Construct a box plot of your data.

_____ Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_____ Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____

_____ Cover sheet: names, class time, and name of your study

_____ Part I: label the sections “Intro” and “Analysis.”

_____ Part II:

_____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.

_____ All graphs requested in the project

_____ All calculations requested to support questions in data

_____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

Note:

Note

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

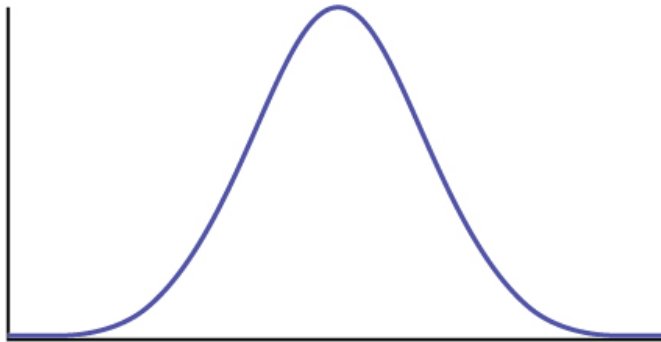
Solution Sheets

Hypothesis Testing with One Sample

Class Time: _____

Name: _____

- a. H_0 : _____
- b. H_a : _____
- c. In words, **CLEARLY** state what your random variable X or P' represents.
- d. State the distribution to use for the test.
- e. What is the test statistic?
- f. What is the p -value? In one or two complete sentences, explain what the p -value means for this problem.
- g. Use the previous information to sketch a picture of this situation. **CLEARLY**, label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
 - i. Alpha: _____
 - ii. Decision: _____
 - iii. Reason for decision: _____
 - iv. Conclusion: _____

- i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

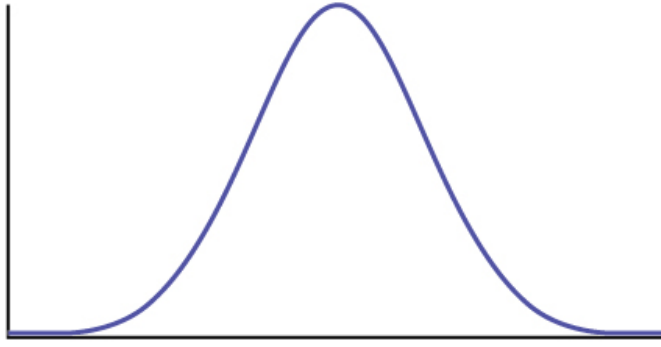


Hypothesis Testing with Two Samples

Class Time: _____

Name: _____

- H_0 : _____
- H_a : _____
- In words, **clearly** state what your random variable $X_1 - X_2$, $P'_1 - P'_2$ or X_d represents.
- State the distribution to use for the test.
- What is the test statistic?
- What is the p -value? In one to two complete sentences, explain what the p -value means for this problem.
- Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

- a. Alpha: _____
- b. Decision: _____
- c. Reason for decision: _____
- d. Conclusion: _____

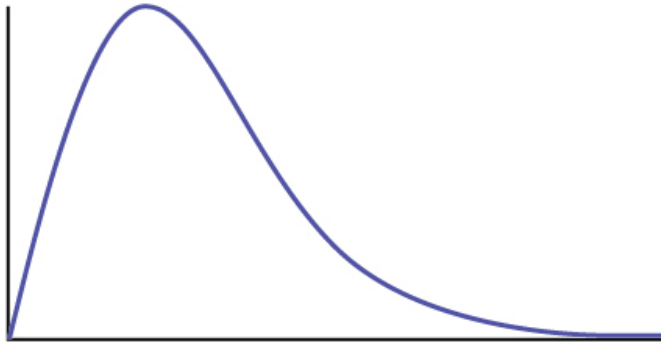
i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: _____

Name: _____

- a. H_0 : _____
- b. H_a : _____
- c. What are the degrees of freedom?
- d. State the distribution to use for the test.
- e. What is the test statistic?
- f. What is the p -value? In one to two complete sentences, explain what the p -value means for this problem.
- g. Use the previous information to sketch a picture of this situation.
Clearly label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.

- i. Alpha: _____
- ii. Decision: _____
- iii. Reason for decision: _____
- iv. Conclusion: _____

F Distribution and One-Way ANOVA

Class Time: _____

Name: _____

- a. H_0 : _____
- b. H_a : _____
- c. $df(n) =$ _____ $df(d) =$ _____
- d. State the distribution to use for the test.
- e. What is the test statistic?
- f. What is the p -value?
- g. Use the previous information to sketch a picture of this situation.
Clearly label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.

- a. Alpha: _____
- b. Decision: _____
- c. Reason for decision: _____
- d. Conclusion: _____

Mathematical Phrases, Symbols, and Formulas

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$

When the English says:	Interpret this as:
There are fewer X than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Formulas

Formula 1: Factorial

$$n! = n(n-1)(n-2)\dots(1)$$

$$0! = 1$$

Formula 2: Combinations

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Formula 3: Binomial Distribution

$$X \sim B(n, p)$$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \text{ for } x = 0, 1, 2, \dots, n$$

Formula 4: Geometric Distribution

$$X \sim G(p)$$

$$P(X = x) = q^{x-1} p, \text{ for } x = 1, 2, 3, \dots$$

Formula 5: Hypergeometric Distribution

$$X \sim H(r, b, n)$$

$$P(X = x) = \left(\frac{\binom{r}{x} \binom{b}{n-x}}{\binom{r+b}{n}} \right)$$

Formula 6: Poisson Distribution

$$X \sim P(\mu)$$

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula 7: Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}, a < x < b$$

Formula 8: Exponential Distribution

$$X \sim \text{Exp}(m)$$

$$f(x) = me^{-mx} \quad m > 0, x \geq 0$$

Formula 9: Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Formula 10: Gamma Function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad z > 0$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(m+1) = m! \text{ for } m, \text{ a nonnegative integer}$$

$$\text{otherwise: } \Gamma(a+1) = a\Gamma(a)$$

Formula 11: Student's t -distribution

$$X \sim t_{df}$$

$$f(x) = \frac{\left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$$Z \sim N(0, 1), Y \sim X_{df}^2, n = \text{degrees of freedom}$$

Formula 12: Chi-Square Distribution

$$X \sim X_{df}^2$$

$$f(x) = \frac{x^{\frac{n-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, x > 0, n = \text{positive integer and degrees of freedom}$$

Formula 13: F Distribution

$$X \sim F_{df(n), df(d)}$$

$$df(n) = \text{degrees of freedom for the numerator}$$

$$df(d) = \text{degrees of freedom for the denominator}$$

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\left(\frac{u}{2}-1\right)} \left[1 + \left(\frac{u}{v}\right)x^{-0.5(u+v)}\right]$$

$$X = \frac{Y_u}{W_v}, Y, W \text{ are chi-square}$$

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\quad}$	The square root of	same
Sampling and Data	π	Pi	3.14159... (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	$Q_3 - Q_1 =$ IQR
Descriptive Statistics	\bar{x}	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean

Chapter (1st used)	Symbol	Spoken	Meaning
Descriptive Statistics	s s_x sX	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x σX	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{ \}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring

Chapter (1st used)	Symbol	Spoken	Meaning
Probability Topics	$P(A B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	\geq	greater than or equal to	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	\leq	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	\neq	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	pdf	prob. density function	same
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	Exp	exponential distribution	same
Continuous Random Variables	k	k	critical value

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	\bar{X}	X -bar	the random variable X - bar
The Central Limit Theorem	μ_x	mean of X	the average of X

Chapter (1st used)	Symbol	Spoken	Meaning
The Central Limit Theorem	$\mu_{\bar{x}}$	mean of X -bar	the average of X -bar
The Central Limit Theorem	σ_x	standard deviation of X	same
The Central Limit Theorem	$\sigma_{\bar{x}}$	standard deviation of X - bar	same
The Central Limit Theorem	ΣX	sum of X	same
The Central Limit Theorem	Σx	sum of x	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same

Chapter (1st used)	Symbol	Spoken	Meaning
Confidence Intervals	t	Student's t - distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with $\alpha/2$ area in right tail	same
Confidence Intervals	p' ; \hat{p}	p -prime; p -hat	sample proportion of success
Confidence Intervals	q' ; \hat{q}	q -prime; q -hat	sample proportion of failure
Hypothesis Testing	H_0	H -naught, H - sub 0	null hypothesis
Hypothesis Testing	H_a	H - a , H -sub a	alternate hypothesis
Hypothesis Testing	H_1	H -1, H -sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error

Chapter (1st used)	Symbol	Spoken	Meaning
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X_1} - \overline{X_2}$	X_1 -bar minus X_2 -bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	μ -1 minus μ -2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	P_1 -prime minus P_2 - prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	p_1 minus p_2	difference in population proportions
Chi-Square Distribution	χ^2	Ky-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency

Chapter (1st used)	Symbol	Spoken	Meaning
Linear Regression and Correlation	$y = a + bx$	y equals a plus b-x	equation of a line
Linear Regression and Correlation	\hat{y}	y-hat	estimated value of y
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ϵ	error	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
Linear Regression and Correlation	1.9s	1.9 times s	cut-off value for outliers




Chapter (1st used)	Symbol	Spoken	Meaning
<i>F</i> - Distribution and ANOVA	<i>F</i>	<i>F</i> -ratio	<i>F</i> -ratio

Symbols and their Meanings

Notes for the TI-83, 83+, 84, 84+ Calculators

Quick Tips

Legend

-  represents a button press
-  represents yellow command or green letter behind a key
-  represents items on the screen

To adjust the contrast

Press



, then hold



to increase the contrast or



to decrease the contrast.

To capitalize letters and words

Press



to get one capital letter, or press



, then



to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake

If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- $4.321 \text{ E } 4 = 4.321 \times 10^4$
- $4.321 \text{ E } -4 = 4.321 \times 10^{-4}$

To transfer programs or equations from one calculator to another:

Both calculators: Insert your respective end of the link cable and press

2nd

, then **[LINK]**.

Calculator receiving information:

Use the arrows to navigate to and select **<RECEIVE>**

Press

ENTER

Calculator sending information:

Press appropriate number or letter.

Use up and down arrows to access the appropriate item.

Press **ENTER** to select item to transfer.

Press right arrow to navigate to and select **<TRANSMIT>**.

Press **ENTER**.

Note:

Note

ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press

2nd

, then **[QUIT]** to exit when done.

Manipulating One-Variable Statistics

Note:

Note

These directions are for entering data with the built-in statistical program.

Data	Frequency
------	-----------

Data	Frequency
-2	10
-1	3
0	4
1	5
3	8

Sample Data We are manipulating one-variable statistics.

To begin:

1. Turn on the calculator.

ON

2. Access statistics mode.

STAT

3. Select **<4:ClrList>** to clear data from lists, if desired.

4

,

ENTER

4. Enter list **[L1]** to be cleared.

2nd

, **[L1]** ,

ENTER

5. Display last instruction.



, 

6. Continue clearing remaining lists in the same fashion, if desired.



,



, ,



7. Access statistics mode.



8. Select 



9. Enter data. Data values go into . (You may need to arrow over to ).

- Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).



,



,



- Continue in the same manner until all data values are entered.

10. In [L2], enter the frequencies for each data value in [L1].

- Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").

4

,

ENTER

- Continue in the same manner until all data values are entered.

11. Access statistics mode.

STAT

12. Navigate to <CALC>.

13. Access <1:1-var Stats>.

ENTER

14. Indicate that the data is in [L1]...

2nd

, [L1] ,

,

15. ...and indicate that the frequencies are in [L2].

2nd

, [L2] ,

ENTER

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

Note:

Note

We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

2nd

, **[STAT PLOT]**

2. Select **<1:plot 1>** to access plotting - first graph.

ENTER

3. Use the arrows navigate go to **<ON>** to turn on Plot 1.

<ON> ,

ENTER

4. Use the arrows to go to the histogram picture and select the histogram.

ENTER

5. Use the arrows to navigate to **<Xlist>**.

6. If "L1" is not selected, select it.



, ,



7. Use the arrows to navigate to .

8. Assign the frequencies to .



, ,



9. Go back to access other graphs.



, 

10. Use the arrows to turn off the remaining plots.

11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.



2. Select each equal sign (=).







3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.



2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.



3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.



2. Select **<9:ZoomStat>**.



3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.



2.
 - $X_{\min} = -2.5$
 - $X_{\max} = 3.5$
 - $X_{scl} = 1$ (width of bars)
 - $Y_{\min} = 0$
 - $Y_{\max} = 10$
 - $Y_{scl} = 1$ (spacing of tick marks on y-axis)
 - $X_{res} = 1$

3. Access graphing mode to see the histogram.

GRAPH

To draw box plots:

1. Access graphing mode.

2nd

, **[STAT PLOT]**

2. Select **<1:Plot 1>** to access the first graph.

ENTER

3. Use the arrows to select **<ON>** and turn on Plot 1.

ENTER

4. Use the arrows to select the box plot picture and enable it.

ENTER

5. Use the arrows to navigate to **<Xlist>**.

6. If "L1" is not selected, select it.

2nd

, **[L1]** ,

ENTER

7. Use the arrows to navigate to **<Freq>**.

8. Indicate that the frequencies are in **[L2]**.

2nd

, **[L2]** ,

ENTER

9. Go back to access other graphs.

2nd

, **[STAT PLOT]**

10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.

11. View the box plot.

GRAPH

, **[STAT PLOT]**

Linear Regression

Sample Data

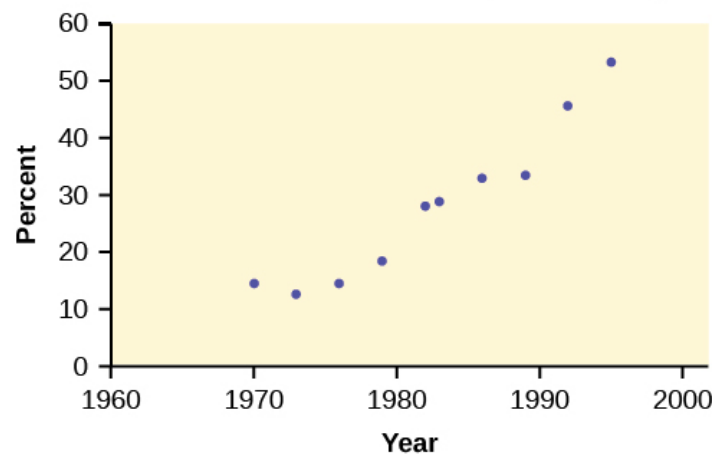
The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16

Year	Student Ethnic Minority Percentage
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage
Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note:

Note

The TI-83 has a built-in linear regression feature, which allows the data to be edited. The x -values will be in **[L1]**; the y -values in **[L2]**.

To enter data and do linear regression:

1. ON Turns calculator on.

ON

2. Before accessing this program, be sure to turn off all plots.

- Access graphing mode.

2nd

, **[STAT PLOT]**

- Turn off all plots.

4

,

ENTER

3. Round to three decimal places. To do so:

- Access the mode menu.

MODE

, **[STAT PLOT]**



- Navigate to **<Float>** and then to the right to **<3>**.

▼

▶

- All numbers will be rounded to three decimal places until changed.



4. Enter statistics mode and clear lists  and , as describe previously.



,



5. Enter editing mode to insert values for x and y .



,



6. Enter each value. Press



to continue.

To display the correlation coefficient:

1. Access the catalog.



, 

2. Arrow down and select 



... ,

ENTER

,

ENTER

3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

STAT

▶

5. Select the form of $y = a + bx$.

8

,

ENTER

The display will show:

LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r = 2\ 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- Percent = $-3176.909 + 1.617$ (year #)

The correlation coefficient $r = 0.961$

To see the scatter plot:

1. Access graphing mode.

2nd

, **[STAT PLOT]**

2. Select **<1:plot 1>** To access plotting - first graph.

ENTER

3. Navigate and select **<ON>** to turn on Plot 1.

<ON>

ENTER

4. Navigate to the first picture.

5. Select the scatter plot.

ENTER

6. Navigate to **<Xlist>**.

7. If **[L1]** is not selected, press

2nd

, **[L1]** to select it.

8. Confirm that the data values are in **[L1]**.

<ON>

ENTER

9. Navigate to **<Ylist>**.

10. Select that the frequencies are in **[L2]**.

2nd

, [L2] ,

ENTER

11. Go back to access other graphs.

2nd

, [STAT PLOT]

12. Use the arrows to turn off the remaining plots.

13. Access window mode to set the graph parameters.

WINDOW

- $X_{\min} = 1970$
- $X_{\max} = 2000$
- $X_{scl} = 10$ (spacing of tick marks on x-axis)
- $Y_{\min} = -0.05$
- $Y_{\max} = 60$
- $Y_{scl} = 10$ (spacing of tick marks on y-axis)
- $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.

15. Press the graph button to see the scatter plot.

GRAPH

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

Y=

2. Access the vars menu and navigate to <5: Statistics>.

VARS

,

3. Navigate to **<EQ>**.
4. **<1: RegEQ>** contains the regression equation which will be entered in Y1.

ENTER

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

GRAPH

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

2nd

, **[LIST]**, **<RESID>**

2. Confirm twice to view the list of residuals. Use the arrows to select them.

ENTER

,

ENTER

3. The critical point for an outlier is: $1.9V \frac{SSE}{n-2}$ where:

- n = number of pairs of data
- SSE = sum of the squared errors
- $\sum \text{residual}^2$

4. Store the residuals in **[L3]**.

STO►

,

2nd

, **[L3]** ,

ENTER

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n - 2 = 8$

2nd

, **[L3]** ,

x²

,

÷

,

8

6. Store this value in **[L4]**.

STO▶

,

2nd

, **[L4]** ,

ENTER

7. Calculate the critical value using the equation above.

1

,

.

,

9

,

×

,

2nd

, [V] ,

2nd

, [LIST]

▶

,

▶

,

5

,

2nd

, [L4] ,

)

,

)

,

ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in [L3] to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of y for various x-values:

There are various ways to determine estimates for "y." One way is to substitute values for "x" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access **DISTR** (for "Distributions").

For technical assistance, visit the Texas Instruments website at <http://www.ti.com> and enter your calculator model into the "search" box.

Binomial Distribution

- **binompdf(n, p, x)** corresponds to $P(X = x)$
- **binomcdf(n, p, x)** corresponds to $P(X \leq x)$
- To see a list of all probabilities for x : 0, 1, . . . , n , leave off the "**x**" parameter.

Poisson Distribution

- **poissonpdf(λ, x)** corresponds to $P(X = x)$
- **poissoncdf(λ, x)** corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value $-1\text{EE}99$ for left bound
- ∞ uses the value $1\text{EE}99$ for right bound

Normal Distribution

- `normalpdf(x, μ, σ)` yields a probability density function value (only useful to plot the normal curve, in which case " x " is the variable)
- `normalcdf(left bound, right bound, μ, σ)` corresponds to $P(\text{left bound} < X < \text{right bound})$
- `normalcdf(left bound, right bound)` corresponds to $P(\text{left bound} < Z < \text{right bound})$ – standard normal
- `invNorm(p, μ, σ)` yields the critical value, k : $P(X < k) = p$
- `invNorm(p)` yields the critical value, k : $P(Z < k) = p$ for the standard normal

Student's t -Distribution

- `tpdf(x, df)` yields the probability density function value (only useful to plot the student- t curve, in which case " x " is the variable)
- `tcdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < t < \text{right bound})$

Chi-square Distribution

- `χ^2 pdf(x, df)` yields the probability density function value (only useful to plot the χ^2 curve, in which case " x " is the variable)
- `χ^2 cdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < X^2 < \text{right bound})$

F Distribution

- `Fpdf($x, dfnum, dfdenom$)` yields the probability density function value (only useful to plot the F curve, in which case " x " is the variable)
- `Fcdf(left bound, right bound, $dfnum, dfdenom$)` corresponds to $P(\text{left bound} < F < \text{right bound})$

Tests and Confidence Intervals

Access **STAT** and **TESTS**.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press **DATA** to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing **STAT** once in the appropriate tests.

Confidence Intervals

- **ZInterval** is the confidence interval for mean when σ is known.
- **TInterval** is the confidence interval for mean when σ is unknown; s estimates σ .
- **1-PropZInt** is the confidence interval for proportion.

Note:

Note

The confidence levels should be given as percents (ex. enter "**95**" or "**.95**" for a 95% confidence level).

Hypothesis Tests

- **Z-Test** is the hypothesis test for single mean when σ is known.
- **T-Test** is the hypothesis test for single mean when σ is unknown; s estimates σ .
- **2-SampZTest** is the hypothesis test for two independent means when both σ 's are known.
- **2-SampTTest** is the hypothesis test for two independent means when both σ 's are unknown.
- **1-PropZTest** is the hypothesis test for single proportion.
- **2-PropZTest** is the hypothesis test for two proportions.
- **χ^2 -Test** is the hypothesis test for independence.

- **χ^2 GOF-Test** is the hypothesis test for goodness-of-fit (TI-84+ only).
- **LinRegTTEST** is the hypothesis test for Linear Regression (TI-84+ only).

Note:

Note

Input the null hypothesis value in the row below "**Inpt.**" For a test of a single mean, " **μ_0** " represents the null hypothesis. For a test of a single proportion, " **p_0** " represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

Tables

The module contains links to government site tables used in statistics.

Note:

Note

When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, January 3, 2009)

- [Student \$t\$ table](#)
- [Normal table](#)
- [Chi-Square table](#)
- [F-table](#)
- All [four tables](#) can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- [95% Critical Values of the Sample Correlation Coefficient](#)